Русская речь
Russian Speech

№ 01 | 2019
С. / Pp. 99–108

Национальный корпус русского языка

Полезные функции в НКРЯ: поиск по части слова и поиск с исключением ненужного элемента

Светлана Олеговна Савчук, Институт русского языка им. В. В. Виноградова РАН (Москва, Россия), savsvetlana@mail.ru

DOI: 10.31857/S013161170003978-4

АННОТАЦИЯ: Национальный корпус русского языка — это не просто большое собрание самых разных текстов, но и разнообразная лингвистическая информация, сопровождающая тексты, и средства поиска информации. Однако многие неподготовленные пользователи не подозревают о богатых возможностях этого лингвистического ресурса и к тому же не любят читать инструкции, поэтому используют его только для простого поиска слов.

Для таких пользователей корпуса мы предлагаем серию заметок, в которых разработчики НКРЯ на примерах конкретных поисковых задач будут знакомить читателей с корпусными инструментами и приемами их использования.

В настоящей заметке рассказывается о средстве, позволяющем осуществлять поиск по части словоформы или лексемы, а также о способе установления фильтров на ненужные единицы, что позволяет получать более точные результаты.

КЛЮЧЕВЫЕ СЛОВА: Национальный корпус русского языка: функциональные возможности, корпусной инструментарий, средства поиска

ДЛЯ ЦИТИРОВАНИЯ: Савчук С. О. Полезные функции в НКРЯ: поиск по части слова и поиск с исключением ненужного элемента // Русская речь. 2019. № 1. С. 99–108. DOI: 10.31857/S013161170003978-4

Russian National Corpus

Useful Functions in Russian National Corpus: search by part of a word and search with the exclusion of an unnecessary element

Svetlana O. Savchuk, Vinogradov Russian Language Institute of the Russian Academy of Sciences (Russia, Moscow), savsvetlana@mail.ru

ABSTRACT: The Russian National Corpus is not only a comprehensive collection of a wide variety of texts, but also a great deal of linguistic information accompanying them, and information search tools. However, many untrained users are unaware of the considerable potential this linguistic resource may offer and, moreover, are unwilling to read the instructions. They tend to use it just for a word search.

For this kind of users of the Corpus, we offer a series of notes, in which the developers, using examples of practical search tasks, will introduce readers to different kinds of corpus tools and techniques to manage them.

This note describes a tool which makes it possible to conduct a search by part of a word form or lexeme, as well as the method of setting filters on unnecessary units and, thereby, getting more accurate results.

KEYWORDS: Russian National Corpus: functionality, corpus tools, search tools

FOR CITATION: Savchuk S. O. Useful Functions in Russian National Corpus: search by part of a word and search with the exclusion of an unnecessary element. Russian Speech = Russkaya Rech'. 2019. No. 1. Pp. 99–108. DOI: 10.31857/S013161170003978-4

Исследователи отмечают, что в современном русском языке функционирует большое количество слов, имеющих в своем составе иноязычные морфемы: **авто**- (*автобус, автобан*), **аэро**- (*аэробус, аэростат*), -**фон** (*телефон, магнитофон, смартфон*), **гипер**- (*гипертекст, гипермаркет, гиперссылка, гиперактивный*), **микро**- (*микромир, микросхема, микробиология*), **супер**- (*супермен, суперкомпьютер, суперфинал)* и т. д. Часть этих слов заимствована из других языков (*автобус, телефон, гипертекст* и др.), часть образована с использованием иноязычных морфем (*автоответчик, суперновость, аэротруба* и др.). С помощью подобных элементов образуется большое количество новых слов, еще не попавших в словари. Можно ли оценить продуктивность той или иной морфемы? В этом может помочь Национальный корпус русского языка благодаря нескольким полезным функциям.

Предположим, нам нужно проверить, активна ли модель с элементом *-навт* (от греч. nautēs — мореплаватель) в современном русском языке. На память сразу приходят слова *космонавт* (космос + nautēs, букв. космоплаватель, ‘человек, совершивший полет в космос’), *астронавт* (астро + nautēs, букв. звездоплаватель, ‘то же, что космонавт’), *аэронавт* (аэро + nautēs, букв. воздухоплаватель), *аргонавты* (Арго + nautēs, ‘древнегреческие герои, совершившие на корабле «Арго» плавание к берегам Колхиды’). Есть ли еще слова на -**навт** в современном русском языке? Поищем ответ в основном корпусе. В этом нам поможет функция «звездочки» (астериска). Звездочка позволяет искать лексемы по какой-то их части, начальной или конечной. Чтобы осуществить запрос, в месте обрыва сегмента (перед ним или после него) надо поставить звездочку (астериск).

На странице лексико-грамматического поиска в поле «слово» записываем ***навт**.

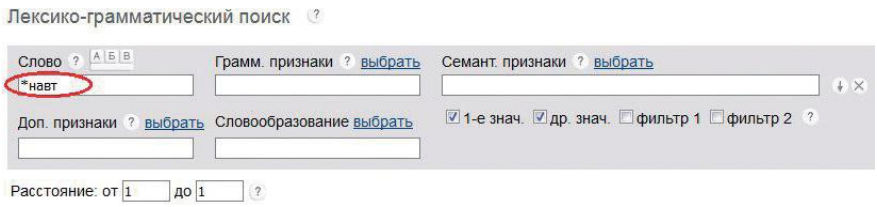


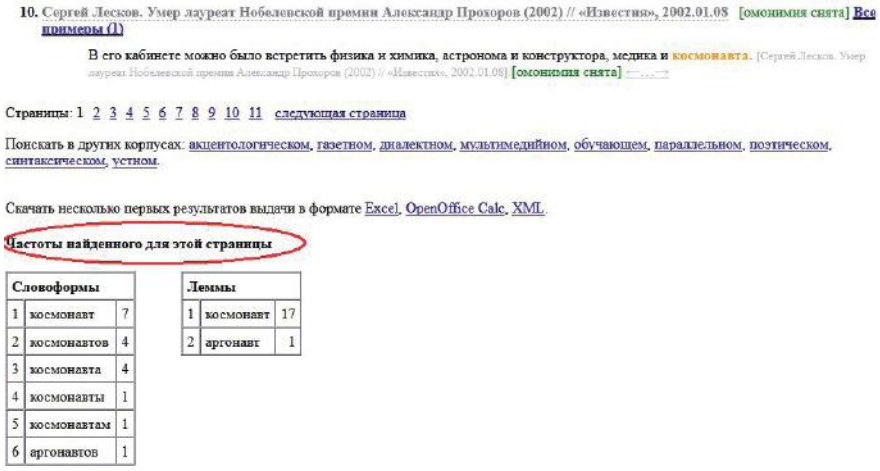
Рис. 1. Запрос для поиска слов, оканчивающихся на *-навт*

Fig. 1. Search query for looking up words ending with *-navt*

В результате такого запроса можно получить около четырех тысяч примеров существительных, оканчивающихся на -**навт**.

- *Выход в открытый космос считается самым сложным заданием для* ***космонавта*** (Известия. 08.01.2003).
- *Сам он уподоблял своё движение от «Мистерии» к «Гармонии» странствию* ***аргонавтов*** *за золотым руном* (Знание — сила. 2003).
- *Я играл* ***космонавта****, а в барокамеру нас снимать не пустили, и мы изображали невесомость в павильонах «Мосфильма»* (Финансовая Россия. 19.09.2002).
- *Но и первоначальные также в их причастности к человеческой истории и вечной с нею борьбе: Орфей в походе* ***аргонавтов*** *усмирял волны, а Пушкин в двух строках* (С. Бочаров. Из истории понимания Пушкина. 1998).
- *Правда, молодой человек был не гусар, не офицер, но в девяностые годы инженер-путеец был фигурой модной, не менее романтической, чем гусар. Нечто вроде* ***космонавта*** *сегодня* (Д. Гранин. Зубр).

Обратите внимание на таблицу внизу каждой страницы. Она называется «Частоты найденного для этой страницы», и в ней указано количество словоформ и лексем (лемм), отвечающих запросу, которые встретились в примерах на данной странице. Таблицы предназначены для быстрого просмотра страниц, поиска нужного слова и оценки его количественных показателей (частотности).



10. Сергей Лесков. Умер лауреат Нобелевской премии Александр Прохоров (2002) // «Известия», 2002.01.08 [омонимия снята] Все примеры (1)

В его кабинете можно было встретить физика и химика, астронома и конструктора, медика и космонавта. [Сергей Лесков. Умер лауреат Нобелевской премии Александр Прохоров (2002) // «Известия», 2002.01.08] [омонимия снята]

Страницы: 1 2 3 4 5 6 7 8 9 10 11 следующая страница

Поискать в других корпусах: акцентологическом, газетном, диалектном, мультимедийном, обучающем, параллельном, поэтическом, синтаксическом, устном.

Скачать несколько первых результатов выдачи в формате Excel, OpenOffice Calc, XML.

Частоты найденного для этой страницы

Словоформы		
1	космонавт	7
2	космонавтов	4
3	космонавта	4
4	космонавты	1
5	космонавтам	1
6	аргонавтов	1

Леммы		
1	космонавт	17
2	аргонавт	1

Рис. 2. Результаты поиска слов, оканчивающихся на *-навт* | **Fig. 2.** Search result for words ending with *-navt*

Большую часть примеров, как мы видим, составляют контексты с существительными *космонавт*, *астронавт* и другими перечисленными выше, что затрудняет поиск менее известных слов. Если исключить из поиска общеизвестные слова, удастся быстрее найти другие слова с элементом **-навт**. Как это сделать? Использовать еще одну полезную функцию — исключение ненужного элемента, для чего служит оператор «минус» перед исключаемой формой или лексемой.

В том же поле «слово» записываем: ***навт -космонавт -астронавт -аэронавт -аргонавт**, что означает «слова, оканчивающиеся на сегмент **-навт**, но не слово *космонавт*, не слово *астронавт*, не слово *аэронавт*, не слово *аргонавт*».

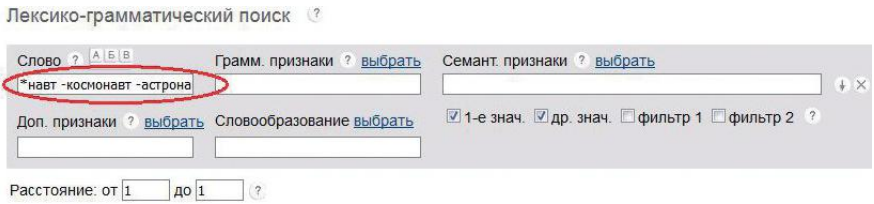


Рис. 3. Запрос для поиска слов, оканчивающихся на *-навт*, за исключением ненужных слов

Fig. 3. Search query for looking up words ending with *-navt*

В результате такого сокращения получаем скромный список из 170 примеров, который легко просмотреть и оценить его состав.

- *В связи с этим невольно складывается впечатление, что НЛО и **НЛО-навты**, если сообщения об их поведении соответствуют действительности, ведут себя именно так, как должны были бы вести себя путешественники в прошлое, прибывшие в наше время из будущих эпох* (В. Комаров. Тайны пространства и времени. 1995–2000).
- *Он был не компьютерщиком, а, как тогда говорили, ученым-**психонавтом*** (В. Пелевин. Любовь к трем цукербринам).
- ***Марсонавты**, которых уже, фактически, вогнали в рабочий режим и в совместную жизнь, приступили к экспериментам* (Детали мира. 2011).
- *Говорил о тех, кого уже три года кряду воспевали газеты: о героях-**стратонавтах*** (Л. Чуковская. Прочерк. 1980–1994).
- *Всего десять лет назад, 15 октября 2003 года, первый китайский **тайконавт**, Ян Ливэй, побывал на околоземной орбите* (Знание — сила. 2013).
- *Ван Юэ — профессиональный психолог, специалист по подготовке и отбору **тэйконавтов*** (Детали мира. 2011).

- *Лишь полное погружение потребует обязательного наличия суперкомпьютера и специальных интерфейсных устройств — «костюмов **виртуанавтов**»* (Воздушно-космическая оборона. 15.08.2003).
- *Раз уж я запустил в 1969 году своего **времянавта** Игоря Одоевцева из 2099 года в пушкинскую эпоху подсмотреть, как дело было, почему бы не подумать о нем сегодня?* (А. Битов. В лужицах была буря… // Звезда. 2002)

Среди найденных примеров есть довольно частотные: *акванавт* (47 вхождений), *стратонавт* (22), *океанавт* (26), *гидронавт* (22), *марсонавт* (6), *тайконавт/тэйконавт* (так называют космонавта в Китае) (4) и др. Это новые слова, заимствования или бывшие неологизмы, которые уже вошли в литературный язык. Другая часть — малочастотные или единичные слова (*бионавт, гелионавт, времянавт, виртуанавт, спелеонавт, НЛО-навты, психонавт, робонавт* и др.), это неологизмы и авторские окказионализмы. Все слова встречаются в текстах, написанных не ранее второй половины XX века, из них почти треть — в текстах после 2000 года, что может свидетельствовать о том, что образованы они по живой продуктивной модели.

Вернемся к звездочке и к иноязычным морфемам. Среди них есть близкие по значению, почти синонимичные, потому что пришли они в русский и другие европейские языки из разных источников: *супер-* (от лат. super ‘над, сверху’) и *гипер-* (от греч. hyper ‘над, сверх’) соотвествуют русской приставке *сверх-*; *контр-* (от лат. contrā ‘против’) и *анти-* (от греч. anti ‘против’) соответствуют русскому *противо-*; *аква-* (от лат. aqua ‘вода’) и *гидро-* (от греч. hydōr ‘вода’) — первая часть сложных слов со значением ‘относящийся к воде’; *авиа-* (от лат. avis ‘птица’) и *аэро-* (от греч. aēr ‘воздух’) — первая часть сложных слов со значением ‘относящийся к авиации’ и др. Интересно проверить, одинаково ли активны эти интернациональные элементы при образовании новых слов в современном русском языке, каким способам отдается предпочтение — с латинскими по происхождению компонентами, греческими или, может быть, их русскими эквивалентами?

Для этого проведем с помощью корпуса предварительное исследование: сравним частоту встречаемости существительных женского рода, оканчивающихся на *-ция* и *-ость*, с разными приставками: **гипер**- (например, *гиперинфляция*) и **супер**- (*суперконцентрация*), а также с русской приставкой **сверх**- (*сверхцентрализация*).

Строим запрос со звездочкой: на странице лексико-грамматического поиска в поле «слово» записываем **гипер*ция**. С помощью минуса исключаем слово *гиперболизация*, в котором не выделяется приставка **гипер**-.

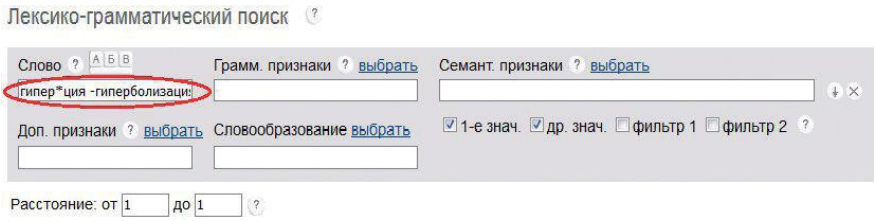


Рис. 4. Запрос для поиска слов, начинающихся на *гипер-* и оканчивающихся на *-ция*

Fig. 4. Search query for looking up words beginning with *giper-* and ending with *-tsiya*

Получаем в ответ около 250 примеров употребления слов заданной структуры.

- *Апологеты долларового стандарта аргументируют его необходимость угрозой* ***гиперинфляции*** *и задачами привлечения иностранных инвестиций* (Эксперт. 2014).
- *Дыхательные практики, то есть* ***гипервентиляция*** *легких, также опасны* (Знание — сила. 2013).
- *Методы лечения* ***гиперфункции*** *щитовидной железы зависят от вызвавшей её причины* (Здоровье. 15.03.1999).
- *В свою очередь эти проблемы тесно связаны с появлением и развитием новых полюсов экономического роста в России, уменьшением* ***гиперцентрализации*** *функций Москвы, организацией нового экономического пространства России с помощью метрополисов (Санкт-Петербург, Нижний Новгород, Казань, Самара, Новосибирск и др.)* (Вопросы статистики. 2004).

Листаем страницы, не забывая просматривать таблицы внизу каждой страницы. Среди найденных слов есть весьма употребительные, встречающиеся по многу раз: *гиперинфляция* (75), *гипервентиляция* (59), *гиперфункция* (16). Есть менее употребительные: *гиперкомпенсация* (8), *гиперпигментация* (5), а есть совсем редкие: *гиперурбанизация* (1), *гиперколонизация* (1). Основная часть примеров — из научных, научно-популярных и публицистических текстов. Все слова книжные, часть из них — научные термины, но самые частотные уже преодолели барьер, отделяющий специальную лексику от общелитературного языка.

Теперь проверим приставочные образования от существительных на *-ость*. По запросу **гипер*ость** получаем меньше примеров — около 150 контекстов.

- *Сейчас наблюдается рост детской расторможенности,* ***гиперактивности****, дефицита внимания, девиантного поведения* (Однако. 2010).
- ***Гиперлокальность*** *— как в джойсовском Дублине — давала Довлатову шанс добраться до основ* (А. Генис. Довлатов и окрестности. 1998).

- *В настоящем региональном исследовании выделяемые смежные регионы разделяются одной* ***гиперплоскостью*** (Вопросы статистики. 18.11.2004).
- *Реальность теперь больше, чем ее любая репрезентация, она —* ***гиперреальность****, то, чему можно найти только сверхчувственный (виртуальный) эквивалент* (В. Подорога. Событие и массмедиа. Некоторые подходы к проблеме. 2010).
- *А это влечет за собой* ***гиперчувствительность*** *зубов, появление кариеса и кучу других проблем* (Твой курс. 2004).

Среди найденных слов много высокочастотных: *гиперчувствительность* (31), *гиперактивность* (22), *гиперплоскость* (8), *гиперреальность* (7) и др. Слова с низкой частотностью относятся скорее не к специальной, а к книжной лексике: *гиперфункциональность* (3), *гипервозбудимость* (3), *гиперлокальность* (1), *гипермобильность* (1), *гиперзанятость* (1), *гиперкритичность* (1), *гиперкультурность* (1) и др.

В целом можно сказать, что с приставкой **гипер**- в современных текстах встречается много существительных женского рода со значением превышения предела, нормы, при этом большая часть из них относится к специальной лексике.

Проделаем ту же работу со словами, содержащими приставку **супер-**. Можно объединить два запроса с помощью оператора «или», который обозначается вертикальной чертой: **супер*ция | супер*ость**.

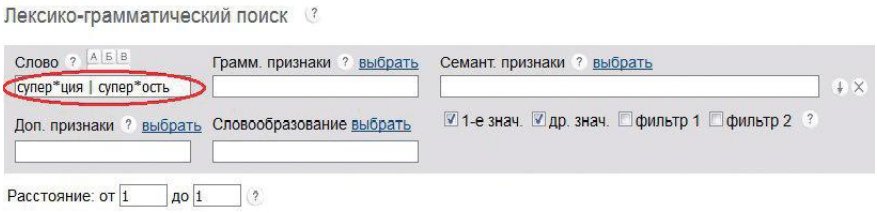


Рис. 5. Запрос для поиска нескольких слов, начинающихся на *супер-* и оканчивающихся на *-ция* или *-ость*

Fig. 5. Search query for looking up words beginning with *super-* and ending with *-tsiya* or *-ost'*

Как это ни удивительно, примеров с приставкой **супер**- гораздо меньше — чуть более 100 на обе словообразовательные модели (нам пришлось исключить многочисленные примеры со словом *суперпозиция*, поскольку в нем приставка не выделяется).

- *Она, возможно, приблизит нас к ответу на вопрос о причинах* ***суперротации*** *венерианской атмосферы* (Энергия: экономика, техника, экология. 1986).

- *Он доминирует в скоростных дисциплинах, да и в* ***суперкомбинации*** *вполне конкурентоспособен* (Эксперт. 2014).
- *А кроме этого, именно к этому празднику «Clearasil» замутил* ***суперакцию*** *— для всех, у кого есть друзья!* (Твой курс. 10.11.2004)
- *Окский пищекомбинат за прошедший год выпустил четыре разновидности водки «Отдохни», сопровождая каждую* ***суперпрезентацией*** (Рекламный мир. 15.02.2000).
- *У него была классная реакция, просто* ***суперреакция*** *супербоксера, он просек все без нудных толкований* (В. Синицына. Муза и генерал. 2002).
- *Никто из полуголодных творцов и не помышлял, что их затягивают в* ***суперпровокацию****, высиженную вот именно в кабинете Килькичева* (В. Аксенов. Таинственная страсть).
- *Людям не хватает силы, отсюда популярность сюжета про* ***суперспособности*** (П. Волошина, Е. Кульков. Маруся).
- *И этот специфический запах просто удушал своей особой* ***суперпроникновенностью*** *и сногсшибаемостью!* (Наука и жизнь. 2008)
- *— Тетя Оля, это ваша дочь, — сказал Гоша торжественно, как будто сообщал маме какую-то* ***суперновость*** (М. Трауб. Плохая мать).
- *На самом деле она вовсе не знала, а только лишь предполагала, что Людмила Панфилова в меру своей* ***суперобщительности*** *и* ***суперкоммуникабельности*** *должна хорошо знать девочку Вику из параллельного класса* (М. Милованов. Естественный отбор).
- *У остальных сотовая суперсвязь: идеальная слышимость в зоне прямой видимости. Все суперпрестижно, все на* ***суперскоростях****! Сверхбыстрый курс языка!* (М. Мишин. Элитарный эксклюзив)
- *Уж за 5 тыр это такая* ***суперокупаемость****, что много что прочее меркнет* (Форум: Бесплатный проезд обойдется в 34,4 млрд. 2012).

Обращает на себя внимание то, что терминов среди существительных совсем немного: *супергравитация* (10), *суперинфекция* (5), *суперротация* (3), *суперкомбинация* (2), *супераддитивность* (1). Гораздо больше примеров использования этой приставки с широким кругом слов нетерминологического характера, в нейтральных и даже бытовых контекстах: *суперцивилизация* (7), *суперкорпорация* (6), *суперакция* (5), *суперпрезентация* (1), *суперпровокация* (1); *суперспособность* (8), *суперновость* (3), *супербодрость* (1), *суперлюбознательность* (1), *супернадежность* (1), *супернелепость* (1), *суперскорость* (1) и др.

Наше небольшое исследование по корпусу наводит на мысль о том, что существование двух приставок с близким значением не случайно. В языке нет ничего бесполезного и избыточного. Близкие по значению приставки *гипер-* и *супер-* как бы поделили сферы влияния: *гипер-* преобла-

дает в научных текстах, по большей части она входит в состав терминов, а приставка *супер-*, сочетаясь как с иноязычными, так и с русскими корнями, активно проявляет себя в других сферах речи — в художественной литературе, в публицистике, бытовой речи, рекламе.

Разумеется, все это только первые наблюдения над материалом. Чтобы убедиться в правильности нашего предположения, потребуется еще много работы: нужно расширить материал, включив в него слова другой словообразовательной структуры (такие, как *гипертекст*, *гиперпространство*, *суперорганизм*, *супертопливо*), тщательно проанализировать контексты с семантической и стилистической точек зрения, обратиться к словарям. И корпус окажет в этом неоценимую помощь.

А как ведет себя в тех же условиях близкая по значению русская приставка **сверх**-? Попробуйте самостоятельно выяснить это с помощью корпуса, сформулировав запросы таким образом, как это было показано в нашей заметке. Просматривать результаты выдачи удобно в формате KWIC (Key Word in Context, ‘ключевое слово в контекстном окружении’ — наиболее распространенный формат представления конкордансов), подробнее об этом мы расскажем в одном из наших следующих очерков.

Источники

Национальный корпус русского языка [Электронный ресурс]. URL: http://ruscorpora.ru

Studiorum. Образовательный портал Национального корпуса русского языка [Электронный ресурс]. URL: https://studiorum-ruscorpora.ru