УДК 811.161.1’373

ОПРЕДЕЛЕНИЕ ФУНКЦИОНАЛЬНОГО ЯДРА РУССКОЙ ЛЕКСИКИ ПО КОРПУСУ ТЕКСТОВ
(НА ПРИМЕРЕ ЧАСТОТНОГО СЛОВАРЯ С. А. ШАРОВА)

Ю. А. Суворова

Воронежский государственный университет

Поступила в редакцию 26 июля 2009 г.

Аннотация: *в статье обсуждается привлечение к выявлению функционального ядра русской лексики данных частотного словаря С. А. Шарова. Устанавливается неприменимость к решению этой задачи Индекса тематической маркированности слова, предложенного А. А. Кретовым. В качестве альтернативы предлагается Индекс функциональной ядерности слова, вычисляемый по формуле: ИнФЯдр = (Q-вес + F-вес) : 2, где Q-вес – вес слова в тексте по частоте, а F-вес – вес слова в тексте по длине. Дается фрагмент функционального ядра русской лексики по словарю-источнику.*
Ключевые слова: *русский язык, параметрический анализ лексики, функциональное ядро лексики, Индекс тематической маркированности, Индекс функциональной ядерности.*

Abstract: *the article explores the possibility of using of S. A. Sharov’s Frequency Dictionary for the extraction of a functional core of Russian vocabulary. The article claims that the Index of Thematic Markedness (InTeM) suggested by A. A. Kretov is not applicable to this task. As an alternative the Index of Functional Nuclearity (InFuN) of a word is used. The Index of Functional Nuclearity is calculated by the following formula: InFuN = = (Q-weight + F-weight)/2, where Q-weight is a word frequency weight in a text, F-weight is a word length weight. The article contains a segment of a Russian functional lexical core extracted from the source dictionary.*
Key words: *russian language, parametric analysis of vocabulary, functional core, Index of Thematic Markedness, Index of Functional Nuclearity.*

Целью данной работы является выявление наиболее активной в функциональном отношении лексики русского языка.

Источником материала для выделения функционального ядра русской лексики выбран частотный словарь С. А. Шарова, взятый из интернета и созданный по корпусу текстов размером 16.336.972 словоформ [1]. Объектом исследования послужил список из 5000 наиболее частотных русских слов. В процессе работы данные словаря обрабатывались в программе Microsoft Excel и по полученным результатам составлялись графики. Основываясь на мнении В. Т. Титова, что ядро лексико-семантической системы составляют существительные, глаголы и прилагательные [2, с. 12; 3, с. 22], мы обращаемся только к этим частям речи.

Функциональный параметр характеризует активность слова в речи (тексте). По мнению В. Т. Титова, одной из важнейших характеристик словаря является длина слова, поскольку частота слов связана обратной зависимостью с их длиной [4, с. 10—11]. Чем употребительнее слово, тем оно короче, и наоборот. Следовательно, длина слова может служить показателем его функциональной активности. При этом буква является наиболее удобной единицей для подсчета длины слова.

По функциональному параметру слова из списка распределились следующим образом (см. таблицу).

Из таблицы видно, что длина слов на основе представленного списка распределяется в промежутке от 1 до 17 букв. Так, по функциональному параметру максимальное значение получили 116 слов длиной в 2 и 3 буквы:

ум, ус, юг, ад, яд, па, як, ля, год, раз, дом, час, мир, вид, ряд, пол, бог, сын, лес, муж, сон, век, зуб, ход, зал, бой, шея, дед, род, тип, суд, дым, шум, нож, вес, лед, бег, имя, шаг, нос, ухо, лоб, бок, сад, рот, дно, еда, миг, бык, пес, поп, дух, гот, эхо, меч, кот, дон, луч, шар, пот, яма, жар, жук, шеф, лук, сэр, дар, суп, жир, тыл, гул, чай, тон, зло, фон, газ, лев, май, ток, зад, цех, луг, вор, рог, люк, вой, рай, хор, акт, маг, мак, сок, гад, рев, щит, рис, раб, сук, мяч, быт, чин, сыр, ген, вал, дуб, мэр, том, кол, мед, бас, лет, бак, око, рок, миф.

Однобуквенными словами являются служебные (предлоги, союзы), поэтому из рассмотрения они исключались, как не несущие лексической семантики.

Самыми длинными являются два слова — *правительственный* и *свидетельствовать*. Они насчитывают в своем составе 17 букв.

Таблица

Распределение первых 5000 наиболее частотных слов по длине

Букв	Слов	Накопл.	Б-вес
1	9	9	1,0000
2	45	54	0,9982
3	157	211	0,9892
4	366	577	0,9578
5	653	1230	0,8846
6	765	1995	0,7540
7	815	2810	0,6010
8	693	3503	0,4380
9	574	4077	0,2994
10	395	4472	0,1846
11	252	4724	0,1056
12	147	4871	0,0552
13	68	4939	0,0258
14	36	4975	0,0122
15	15	4990	0,0050
16	8	4998	0,0020
17	2	5000	0,0004

Распределение длины слова в русском языке представлено на рис. 1.

На рис. 1 четко видно, что наибольшее количество слов имеет в своем составе 7 букв, а слова, находящиеся рядом с ними, насчитывают 6 и 8 букв.

А. А. Кретов предлагает совместить в исследовании не только длину слов, но и частоту их употребления [5]. Для вычисления частотного веса используется та же формула, что и для буквенного (функционального) веса. Вычитая из частотного веса каждого слова его буквенный вес, мы можем определить Индекс тематической маркированности слова (ИнТеМ). Результатом этой операции является показатель, который с положительным значением веса указывает на то, что носители языка употребляют эти слова чаще, чем предписывает их длина, а отрицательное значение индекса тематической маркированности указывает на то, что носители языка используют эти слова реже, чем позволяет их длина.

На рис. 2 представлена стратификация 5000 наиболее употребительных слов русского языка по индексу тематической маркированности. По вертикали указывается индекс тематической маркированности, а по горизонтали — номер слова в порядке убывания индекса тематической маркированности.

Итак, по индексу тематической маркированности наиболее часто употребляемыми в русском языке оказались следующие слова:

рассказывать 0,868, остановиться 0,867, почувствовать 0,861, государственный 0,859, единственный 0,853, представлять 0,842, человеческий 0,841, человеческий 0,809, существовать 0,84, чувствовать 0,833, разговаривать 0,83, собственный 0,828, возвращаться 0,824, становиться 0,817, политический 0,813, возможность 0,809.

Указанные слова насчитывают в своем составе от 15 до 11 букв.

Наименьшим значением по индексу тематической маркированности обладают слова, насчитывающие в своем составе от 2 до 4 букв:

миф 0,969, рок 0,963, овощ 0,954, око 0,954, ля 0,952, матч 0,951, мисс 0,948, вить 0,944, сейф 0,938, сало 0,935, роща 0,935, грек 0,935, бак 0,9344, язва 0,932,

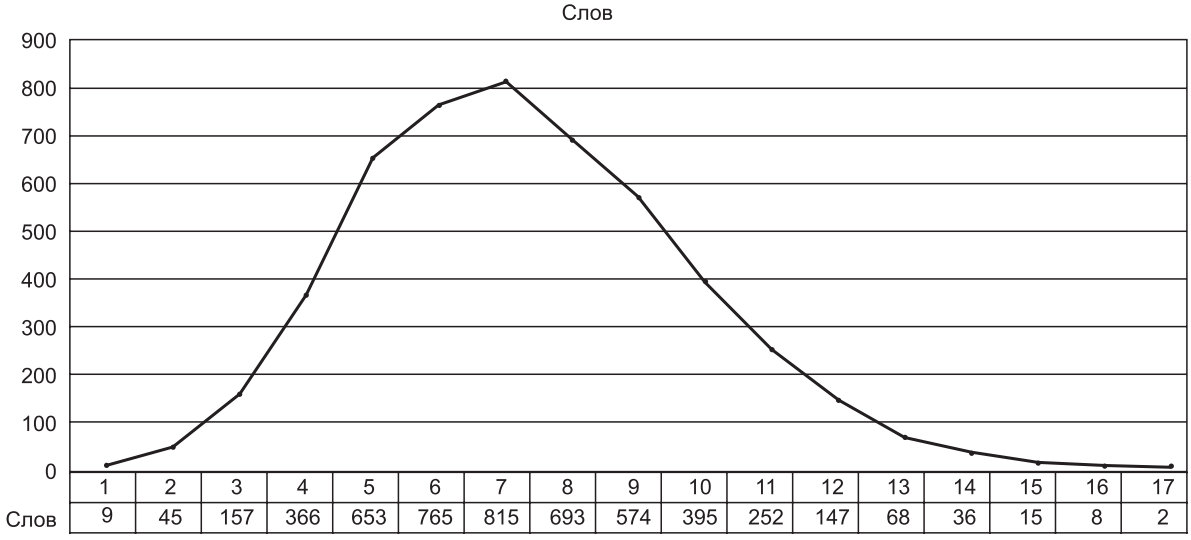

Рис. 1. Распределение по длине 5000 наиболее употребительных русских слов

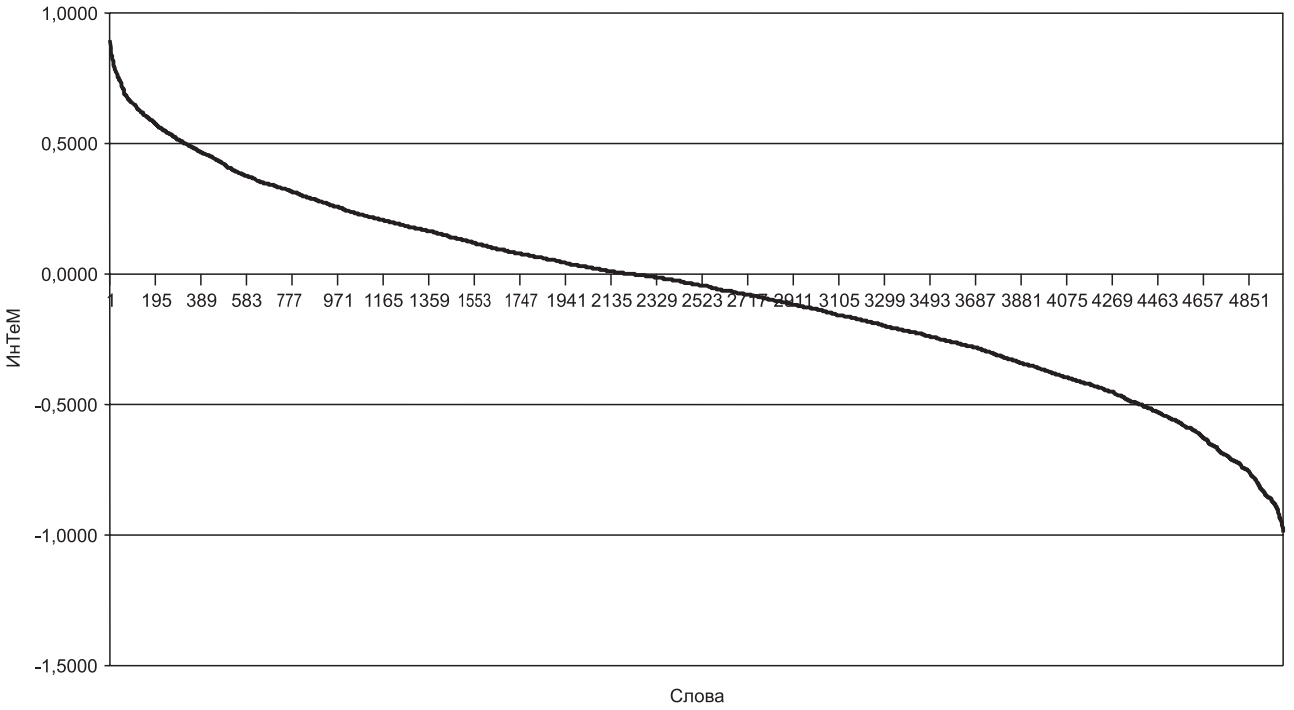

Рис. 2. Стратификация 5000 наиболее употребительных слов русского языка по индексу тематической маркированности

удав 0,932, тигр 0,92, лет 0,914, бас 0,911, стон 0,909, мед 0,907, семя 0,903, як 0,901, эфир 0,9.

Как видим, слова с минимальным значением ИнТеМа столь же мало похожи на ядерную лексику русского языка, как и слова с максимальным ИнТеМом.

Попробуем поискать ядерную лексику вблизи нулевого значения ИнТеМа. Для этого возьмем значения ИнТеМа по модулю и отсортируем их в порядке возрастания. Получаем следующий результат: *автор 0,0002, подтверждать 0,0004, переглянуться 0,0004, вселенная 0,0004, килограмм 0,0004, окно 0,0006, мужик 0,0006, стыдный 0,0006, непривычный 0,0006, перестройка 0,0006, частный 0,0006, похоронить 0,0008, год 0,001, мать 0,001, желтый 0,001, роскошный 0,001, схватиться 0,001, повесть 0,0012, тарелка 0,0012, бывший 0,0014, свет 0,0014, приказывать 0,0016, убить 0,0018, громкий 0,0018, глянуть 0,0018…*

Вряд ли слова *автор, стыдный, перестройка, частный, повесть, тарелка* относятся к ядру русского языка, а между тем именно они попали в первую сотню.

В связи с этим мы делаем вывод, что ИнТеМ, предложенный А. А. Кретовым для выделения тематически нейтральной и маркированной лексики, не может быть использован для выделения функционального ядра лексики.

Однако сама идея одновременного учета частотных весов слов и их длины должна быть сохранена и удержана.

Вместо Индекса тематической маркированности слова (ИнТеМ), вычисляемого по формуле ИнТеМ = Q-вес – F-вес, где Q-вес – вес слова по частоте, F-вес – вес слова по длине, мы предлагаем использовать Индекс функциональной ядерности (ИнФЯдр), вычисляемый по формуле:

$$\text{ИнФЯдр} = (\text{Q-вес} + \text{F-вес}) : 2,$$

где Q-вес – вес слова по частоте; F-вес – вес слова по длине.

Данный индекс суммирует два функциональных ядра: *спонтанное* (выделяемое по Q-весу) и *стабильное* (выделяемое по F-весу) и усредняет их показания деление на 2, позволяющим ИнФЯдр оставаться в пределах единицы.

Если мы отсортируем те же 5000 слов частотного словаря С. А. Шарова по ИнФЯдру, то получим весьма правдоподобное функциональное ядро русской лексики (см. приложение).

Полученный результат представляется нам положительным, что позволяет рекомендовать выделение

функционального ядра с помощью ИнФЯдра при работе с большими корпусами текстов.

Следует отметить, что результаты являются несколько огрубленными, поскольку данные словаря учитывают лишь частоту слова в его словарной форме (лемме). Перспективой исследования является учет частотности словоформ, например, по Национальному корпусу русского языка (см. www.ruscorpora.ru).

ЛИТЕРАТУРА

1. *Шаров С. А*. Частотный словарь русского языка / С. А. Шаров. URL: http://www.artint.ru/projects/frqlist.asp

2. *Титов В. Т*. Частная квантитативная лексикология романских языков / В. Т. Титов. — Воронеж: Изд-во Воронеж. гос. ун-та, 2004. — 552 с.

3. *Титов В. Т*. Методические указания по выделению параметрического ядра лексики (для филологов: студентов и аспирантов) / В. Т. Титов. — Воронеж, 2006. — 52 с.

4. *Титов В. Т*. Общая квантитативная лексикология романских языков / В. Т. Титов. — Воронеж: Изд-во Воронеж. гос. ун-та, 2002. — 240 с.

5. *Кретов А. А*. Метод формального выделения тематически нейтральной лексики : (на примере старославянских текстов) / А. А. Кретов // Вестник Воронежского государственного университета. Сер.: Системный анализ и информационные технологии. — 2007 — № 1. — С. 81—90.

Приложение

Функциональное ядро русской лексики, полученное с помощью ИнФЯдра

год 0,99, раз 0,987, дом 0,985, час 0,98, быть 0,978, мир 0,978, вид 0,977, ряд 0,977, мочь 0,976, рука 0,973, дело 0,972, есть 0,972, пол 0,971, глаз 0,971, день 0,971, бог 0,971, идти 0,97, друг 0,969, лицо 0,968, сын 0,968, лес 0,967, жить 0,967, нога 0,967, вода 0,964, стол 0,963, сила 0,963, отец 0,963, ночь 0,962, дать 0,961, жена 0,959, свет 0,959, мать 0,958, окно 0,958, душа 0,956, утро 0,956, муж 0,954, сон 0,953, уйти 0,953, пора 0,952, век 0,951, тело 0,951, мама 0,95, зуб 0,95, путь 0,95. язык 0,948, небо 0,947, ход 0,946, ум 0,944, зал 0,941, труд 0,939, снег 0,937, знать 0,937, время 0,937, двор 0,937, стать 0,936, угол 0,936, губа 0,935, дядя 0,935, жизнь 0,935, рота 0,933, море 0,933, врач 0,932, край 0,932, слово 0,932, река 0,932, место 0,932, мера 0,932, бой 0,932, шея 0,932, дед 0,931, новый 0,931, род 0,931, вещь 0,931, боль 0,93, иметь 0,93, взять 0,929, пить 0,929, дверь 0,929, пара 0,928, речь 0,928, земля 0,928, конец 0,928, голос 0,928, город 0,928, пойти 0,926, цель 0,926, выйти 0,925, след 0,925, война 0,924, белый 0,923, тип 0,922, суд 0,922, вера 0,922, удар 0,922, ждать 0,922, найти 0,921, часть 0,921, игра 0,921, книга 0,92. улица 0,92, дым 0,92, вечер 0,919, народ 0,919, плечо 0,919, счет 0,919, мысль 0,917, общий 0,917, поэт 0,915, кожа 0,914, стена 0,914, лист 0,914, право 0,914, роль 0,913, рост 0,913, месяц 0,913, брать 0,913, целый 0,912, спина 0,912, баба 0,912, вести 0,911, дочь 0,911, живой 0,911, шум 0,91, нож 0,909, союз 0,908, лето 0,908, цвет 0,908, школа 0,907, кровь 0,907, мозг 0,906, воля 0,905, дама 0,905, спать 0,905, войти 0,904, берег 0,904, факт 0,903, семья 0,903, рыба 0,903, вино 0,903, ответ 0,902, зима 0,901, смысл 0,901, смочь 0,901, форма 0,90, щека 0,90, грудь 0,899, ехать 0,899, немец 0,899, знак 0,899, огонь 0,898, армия 0,898, тема 0,897, гость 0,897, ветер 0,897, курс 0,897, закон 0,896, срок 0,895, ужас 0,895, звать 0,895, крик 0,895, очко 0,894, номер 0,893, глава 0,893, серый 0,893, страх 0,892, снять 0,892, ключ 0,892, черт 0,892, пыль 0,891, связь 0,891, даль 0,89, класс 0,89, знать 0,889, худой 0,889, вес 0,889, число 0,886, рубль 0,886, мужик 0,885, автор 0,885, лед 0,884, убить 0,884, пиво 0,884, мясо 0,883, наука 0,882, кухня 0,882, роман 0,881, рада 0,881, кино 0,88, волк 0,879, трава 0,879, бег 0,879, вход 0,878, слать 0,877, урок 0,877, имя 0,877, точка 0,876, долг 0,876, шаг 0,876, толпа 0,876, нос 0,876, ухо 0,875, толк 0,875, тихий 0,875, сапог 0,875, дождь 0,874, слой 0,873, ворот 0,873, шкаф 0,873, сосед 0,873, изба 0,873, лапа 0,872, выход 0,872, совет 0,872, дурак 0,872, левый 0,87, куча 0,87, орган 0,869, нести 0,869, хотеть 0,868, пища 0,868, милый 0,868, голова 0,868, видеть 0,868, стоять 0,867, лоб 0,867, думать 0,867, повод 0,865, птица 0,865, сухой 0,865, сидеть 0,865, понять 0,865, бок 0,864, делать 0,864, сад 0,864, март 0,864, сеть 0,863, нужный 0,863, работа 0,863, фронт 0,863, район 0,863, итог 0,863, мешок 0,862, завод 0,862, давать 0,862, кулак 0,862, рот 0,861, кусок 0,861, машина 0,86, брат 0,86, умный 0,86, случай 0,86, малый 0,86, любить 0,859, старый 0,859, начало 0,859, вопрос 0,859, узкий 0,858, гора 0,858, деньги 0,858, пауза 0,857, минута 0,857, правда 0,857, страна 0,857, дно 0,856, хвост 0,856, дорога 0,856, лежать 0,856, стихи 0,856, хлеб 0,856, еда 0,856, писать 0,855, враг 0,855, вагон 0,855, решить 0,855, метр 0,854, масса 0,854, голый 0,854, идея 0,854, звук 0,854, отдел 0,853, сцена 0,853, билет 0,853, взгляд 0,853, ходить 0,853, петь 0,853, бывать 0,852, полный 0,852, запад 0,852, член 0,852, россия 0,852, узнать 0,852, ящик 0,851, этаж 0,851, стул 0,85, яркий 0,85, миг 0,85, учить 0,85, тень 0,85, поле 0,85.

Воронежский государственный университет
Суворова Ю. А., преподаватель кафедры теоретической и прикладной лингвистики
E-mail: SuvorovaJ84@mail.ru
Тел.: 8-903-655-09-93

Voronezh State University
Suvorova Ju. A., Lecturer of the Theoretical and Applied Linguistics Department
E-mail: SuvorovaJ84@mail.ru
Tel.: 8-903-655-09-93