КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 001.103:002

«БЕЛЫЙ ДОМ» И «ЧЁРНАЯ ДЫРА»: АЛГОРИТМ ОПРЕДЕЛЕНИЯ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ СЛОВ ПО ИХ СОЧЕТАЕМОСТИ

А. А. Кретов

Воронежский государственный университет

Поступила в редакцию 28.11.2016 г.

Аннотация. На примере сочетаемость в НКРЯ слов *белый* и *черный* предъявляется Алгоритм Формального Определения Семантической Близости Слов (АФОСеБС). Вычисляется семантическая близость слов. Выявляется их специфическая сочетаемость.
Ключевые слова: компьютерная лингвистика, лексическая семантика, семантическая близость слов, сочетаемость слов, алгоритм формального определения семантической близости слов, формальное выделение идиом.
Annotation. The article focuses upon algorithm of formal word semantic proximity assessment (AFoWSPA). The results of the computational experiment are discussed.
Keywords: computational linguistics, lexical semantics, semantic proximity of the words, algorithm of the formal word semantic proximity assessment, formal assessment of idioms.

Данная статья является продолжением разработки проблематики, намеченной ранее в публикации [Воронина 2010].

Предлагаемый нами *Алгоритм Формального Определения Семантической Близости Слов (АФОСеБС)* основан на постулате: семантика двух слов тем ближе, чем больше общего в их дистрибуции (сочетаемости). При этом уточним, что при формальном подходе к анализу текста, предлагаемом нами, под **словом** мы имеем в виду парадигматическое слово в понимании А. А. Зализняка [Зализняк 2002:20], а под **лексической сочетаемостью** слова L понимается множество слов, находящихся в позиции P по отношению к слову L в обследованном корпусе текстов и обладающих определённым **свойством** (в нашем случае при L = прилагательное это свойство – «быть существительным»).

Для проверки работы алгоритма и его демонстрации используется простейший случай – исследуется близость сочетаемости двух прилагательных: *белый* (слово L_1) и *чёрный* (слово L_2). Из Национального корпуса русского языка (http://www.ruscorpora.ru/search-main.html; далее – НКРЯ) извлекаются *употребления* (последовательности символов между показателями конца предложения[1], содержащие L) данных слов: 3082 для *белый* и 3109 для *чёрный*.

Для русского языка при исследовании сочетаемости прилагательных позицией P является позиция первого слова справа от слова L. Из полученного множества исключаются все слова, не являющиеся существительными: служебные слова, прилагательные, глаголы, наречия и т. д.

[1]Показателями конца предложения считаются сочетания символов: точка+пробел, восклицательный знак+пробел, вопросительный знак+пробел, многоточие+пробел (последнее сочетание встречается и внутри предложения).

Словоформы существительных *лемматизируются* (приводятся к словарной форме – *лемме*), и для каждого парадигматического слова (представленного в тексте конкретным сегментом [Зализняк 2002:19]) вычисляется его частота как сумма частот входящих в его парадигму абстрактных сегментов [Зализняк 2002:19].

Общая сумма абсолютных частот всех субстантивных словоформ, встречающихся справа от данного прилагательного принимаем за 1 (или за 100 %).

Для каждой леммы вычисляем её относительную частоту. Поскольку прилагательных у нас два, множеств лемматизированных существительных – тоже два. Соответственно, если какое-то из лемматизированных слов входит в оба множества, у него будет две относительных частоты.

Справа от прилагательного *белый* в выборке НКРЯ встретилось 799 разных существительных 2092 раза. Обозначим это как размерность множества Б = 2092.

Справа от прилагательного *черный* в выборке НКРЯ встретилось 776 разных существительных 2157 раз. Размерность множества Ч = 2157.

Полученные данные позволяют нам оценить соотношение информации и шума в выборках из НКРЯ. Для прилагательного *белый* информативность выборки составляет 2092/3082 = 68 %, а для прилагательного *чёрный* – 2157/3109 = 69 %. В первом случае на долю шума приходится 32 %, во втором – 31%. Несколько огрубляя, можно сказать, что в рамках обследованного материала и при изложенном подходе данные НКРЯ содержат $\frac{2}{3}$ информации и $\frac{1}{3}$ шума. Это знание представляется полезным при оценке точности получаемых данных.

Теперь определим зону пересечения множеств Б и Ч. Назовём её множеством БЧ, в которое вошло 261 слово: *автомобиль, аист, акула, бабочка, балахон, башенка, бисер, блок, борода, бородка, брови, брюки, буква, бумага, бусинка, верх, ветка, взгляд, вино, вода, водолазка, Волга, волна, волосы, ворона, воротник, гарус, глаза, голова, голубь, город, горошек, гребешок, гроб, грудь, двор, день, джинсы, диск, дом, дуб, дым, дымок, жидкость, зависть, занавеска, зарплата, заяц, звездочка, земля, зерно, зима, змея, золото, зонтик, зубы, изба, камень, капля, квадрат, китель, клеенка, клетка, клобук, клочок, клубок, кожа, колготки, колено, колонна, колпак, кольцо, комбинезон, комната, конверт, конь, кора, коридор, король, коса, кость, костюм, костюмчик, косынка, кот, кофточка, кошка, крапина, крапинка, краска, крест, кролик, круг, крыло, кудри, кулак, курица, куртка, ладонь, лапа, лебедь, лес, линия, лист, лицо, локон, лошадь, луна, люди, маечка, майка, малина, мантия, маска, масса, масть, мгла, мех, мешок, мир, молния, море, мох, мрамор, мундир, муть, муха, мышь, надпись, народ, население, небо, нитка, ниточка, нить, ноготь, нос, носки, ночь, облако, огонь, одежда, одеяние, окраска, оправа, очи, пакет, пальто, пар, парень, парус, паук, пена, передник, переплет, перец, перчатка, песок, пешка, пиар, пигмент, платок, платочек, платье, плащ, плита, поверхность, повязка, покрывало, поле, полоса, полоска, потолок, пояс, пространство, прядь, пряжа, птица, пуговка, пудель, пузырек, пыль, пятно, пятнышко, резинка, ресницы, решетка, роллс-ройс, рот, рубаха, рубашка, рука, рыба, саксаул, свет, сила, символ, скала, след, смерть, смородина, снаряд, снег, спина, список, ствол, стена, стол, сторона, страница, тапочки, тарелка, тело, тесто, тесьма, тигр, ткань, тон, точка, трава, трико, трубка, трюфель, туман, тюльпан, угол, улица, усы, ухо, фартук, фигура, фигурка, флаг, фон, форма, футболка, халат, хаммер, хлеб, царство, цвет, циферблат, челка, человек, шаман, шапка, шапочка, шар, шевелюра, шлем, шнурок, шрам, штаны, штрих, шуба, щека, щенок, экран, юбка, ядро, ящик*.

При этом дистрибуция прилагательного *белый* имеет 32,7 % общих с прилагательным *чёрный* существительных, а *черный* – 33,6 % существительных общих с прилагательным *белый*. Таким образом, можно заключить, что в своем качественном аспекте дистрибуция прилагательных *белый* и *чёрный* совпадает на $\frac{1}{3}$.

Относительная частота лемм множества Б, входящих в подмножество БЧ, равна 56,17 %,

тогда как относительная частота лемм множества Ч, входящих в подмножество БЧ, равна 51,14 %.

Теперь у нас есть всё необходимое для вычисления силы связи между прилагательными *белый* и *чёрный*.

$$\text{ИСеБлиС}^{\text{белый}}{}_{+\text{чёрный}} = \\ = \frac{56{,}17\,\% + 51{,}14\,\%}{200\,\%} = 53{,}66\,\%. \quad (1)$$

Относительную частоту элементов множеств Б и Ч мы можем использовать также для выявления специфической (маркированной) или нейтральной для прилагательных *белый-чёрный* сочетаемости.

Для этого предпочтительнее воспользоваться континуальной шкалой, используя разность относительных частот лемм, входящих в множества Б и Ч:

$$\text{Разн}^{\text{Б}}{}_{\text{ч}}\text{отн} = \\ = \text{ОтнЧаст} - \text{л}_i^{\text{Б}} - \text{ОтнЧаст} - \text{л}_i^{\text{ч}}. \quad (2)$$

Леммы с максимально положительной разностью будут специфичными для прилагательного *белый*, леммы с максимальной отрицательной разностью будут специфичны для прилагательного *черный*, а леммы со значениями разности относительных весов близкой к 0 будут указывать на то, что сближает и объединяет множества Б-Ч, а значит, и семантику слов, сочетающихся с леммами *белый* и *черный*.

На Рис. 1 видно, что большая часть лемм нейтральна по отношению к прилагательным *белый-чёрный*, но множества маркированных лемм выражены весьма ярко.

К положительно маркированным леммам типичным для сочетаемости прилагательной белый (в порядке убывания специфики) относятся 106 существительных: *дом – 3,87 %; свет – 2,92 %; халат – 2,58 %; рубашка – 1,73 %; пятно – 1,17 %; карлик – 0,96 %; гриб – 0,91 %; ночь – 0,78 %; бумага – 0,72 %; медведь – 0,72 %; зубы – 0,67 %; простыня – 0,67 %; движение – 0,57 % и др.*

К отрицательно маркированным леммам типичным для сочетаемости прилагательной *чёрный* (в порядке убывания специфики) относятся 124 существительных: *дыра –5,10 %; море –2,73 %; глаза –1,61 %; смородина –1,34 %; дерево –1,07 %; список –0,88 %; тень –0,88 %; ящик –0,83 %; юмор –0,83 %; перец –0,74 %; рынок –0,65 %; икра –0,60 %; золото –0,59 %; точка –0,55 % и др.*

***Белый дом* – прежде всего идиома:**

“Здание

• Белый дом (англ. The *White House*) – здание резиденции президента США.

• Белый дом – здание правительства Российской Федерации – России.

• Белый дом – здание правительства Свердловской области.

• Белый дом – здание усадьбы горнозаводчиков Демидовых.

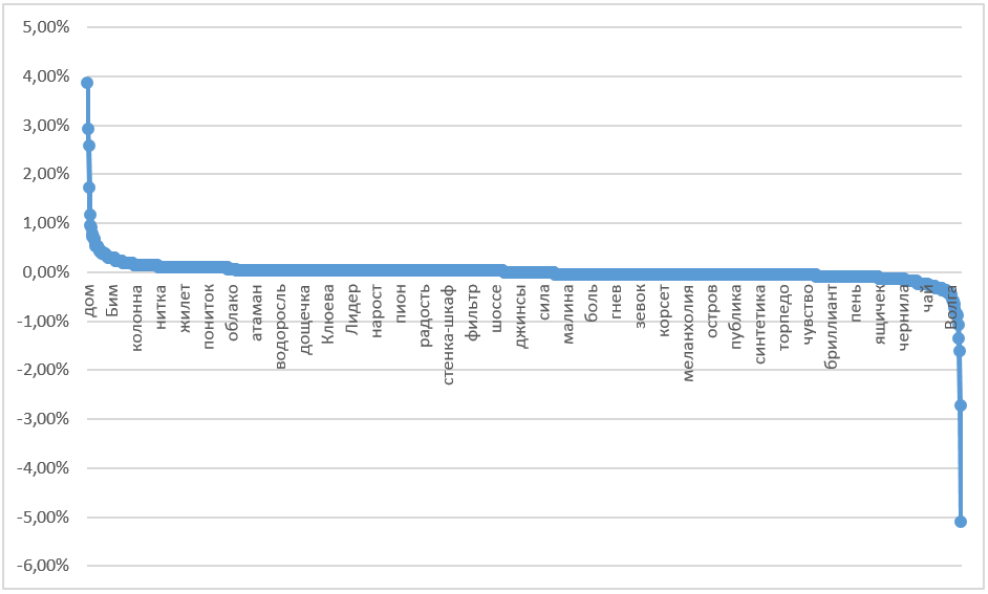

Рис. 1. Распределение лемм по разности их относительных частот во множествах Б и Ч: Ботн-Чотн

• Белый дом – здание правительства Киргизии.

• Белый дом – здание парламента Абхазии, в городе Сухум.

• Белый дом – здание, где жил и работал генерал-губернатор Восточной Сибири Муравьёв-Амурский.

• Белый дом – средневековый водный замок, расположенный в городе Кёльн (Северный Рейн-Вестфалия, Германия).

Иное

• Белый дом – скальный дворец индейцев-пуэбло, сохранившийся в национальном парке Каньон Де Шейи.

• «Белый дом» (2006) – фильм длительностью 690 минут (11 часов), один из самых длинных в мире.

• Белый дом – село в Сарыагашском районе Южно-Казахстанской области Казахстана.

• Белый дом ночью – картина нидерландского живописца Винсента ван Гога. [https://ru.wikipedia.org/wiki/%D0%91%D0%B5%D0%BB%D1%8B%D0%B9_%D0%B4%D0%BE%D0%BC_(%D0%B7%D0%BD%D0%B0%D1%87%D0%B5%D0%BD%D0%B8%D1%8F)].

Чёрная дыра – также устойчивое словосочетание с единым смыслом:

• **Чёрная дыра** – область в пространстве-времени, гравитационное притяжение которой настолько велико, что покинуть её не могут даже объекты, движущиеся со скоростью света.

• **Чёрная дыра (фильм, 1979)** (англ. *The Black Hole*) – фантастический фильм, США, 1979 год. Режиссёр – Гэри Нелсон.

• **Чёрная дыра (фильм, 2000)** (англ. *Pitch Black* – «Кромешная тьма») – фантастический фильм, США, 2000 год. Режиссёр – Дэвид Туи.

• **Чёрная дыра (фильм, 2006)** (англ. *The Black Hole*) – фантастический фильм, США, 2006 год. Режиссёр – Тибор Такач.

• **Чёрная дыра (Сверхъестественное)** (англ. *Mystery Spot*) – 11 эпизод 3 сезона телесериала «Сверхъестественное».

• **Чёрная дыра (комикс)** – серия комиксов из 12 выпусков, написанная Чарльзом Бёрнсом, 1995 – 2005 годы. [https://ru.wikipedia.org/wiki/%D0%A7%D1%91%D1%80%D0%BD%D0%B0%D1%8F_%D0%B4%D1%8B%D1%80%D0%B0_(%D0%B7%D0%BD%D0%B0%D1%87%D0%B5%D0%BD%.

Как видим, типичная (маркированная) сочетаемость свойственна прежде всего идиомам – устойчивым словосочетаниям с единым смыслом. Этот попутный результат открывает путь к формализации и шкалированию понятия идиоматичности и компьютерному выделению идиом в текстах.

СПИСОК ЛИТЕРАТУРЫ

1. Алгоритмы определения семантической близости ключевых слов по их окружению в тексте / И. Е. Воронина, А. А. Кретов, И. В. Попова // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2010. – № 1. – С. 148–153.

2. *Зализняк А. А.* «Русское именное словоизменение» с приложением работ по современному русскому языку и общему языкознанию. – М. : Языки славянской культуры, 2002. – I-VIII, 752 с. – (Studia philological).

Кретов Алексей Александрович – д-р филол. наук, проф., зав. каф. теоретической и прикладной лингвистики, Воронежский государственный университет.
Тел.: (+7 4732) 204-149
E-mail: tipl@rgph. vsu.ru

Kretov A. A. – Doctor of Philology Sciences, Professor, the Head of the dept. of Theoretical and Applied Lingvistics, Voronezh State University.
Tel.: (+7 4732) 204-149
E-mail: tipl@rgph.vsu.ru