УДК 519.7 к.т.н., доц. **Хайрова Н.Ф.** (НТУ «ХПИ»)

КОНЦЕПТУАЛЬНАЯ СХЕМА ИДЕНТИФИКАЦИИ СМЫСЛА ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ

В статті пропонується концептуальна схема ідентифікації смислу елементів ієрархічної мовної системи, яка здійснюється за рахунок факторизації простору концептів, представлених знаками лінгвістичних смислових одиниць. Доводиться можливість використання даної концептуальної схеми для здійснення семантичної класифікації елементів різних рівнів мовної системи. Пропонується використовувати підходи і методи теорії інтелекту для реалізації даної схеми в формальних моделях семантичної обробки на мові логіки предикатів.

Ключові слова: ідентифікація знань, мовна система, лінгвістичні одиниці, теорія інтелекту, мова логіки предикатів.

В статье предлагается концептуальная схема идентификации смысла элементов иерархической языковой системы, осуществляемая за счет факторизации пространства концептов, выражаемых знаками лингвистических смысловых единиц. Доказывается возможность использования данной концептуальной схемы для осуществления семантической классификации элементов различных уровней языковой системы. Предлагается использовать подходы и методы теории интеллекта для реализации данной схемы в формальных моделях семантической обработки на языке логики предикатов.

Ключевые слова: идентификация знаний, языковая система, лингвистические единицы, теория интеллекта, язык логики предикатов.

The article proposes conceptual framework for identifying the hierarchical sense elements of the language system, implemented by the factorization of the space of concepts, expressed signs of language semantic units. It is proved the possibility of using of the conceptual framework for semantic classification of the elements of different levels of the language system. It is proposed to use the approaches and methods of the artificial intelligence for the implementation of this scheme in the formal models of semantic processing in the predicate logic language.

Keywords: identification of knowledge, language system, linguistic units, theory of artificial intelligence, the language of predicate logic.

Актуальность исследования. К настоящему моменту различными хранилищами знаний накоплены огромные информационные массивы, однако отсутствие возможности оперативно получить наиболее актуальную и полную информацию по конкретной теме обесценивает значительную часть накопленных информационных ресурсов. Поэтому все больше научных исследований фокусируется на разработке формальных моделей и методов обработки естественного языка. Исследования в области автоматической обработки текстов в Европе и США привлекают внимание крупнейших частных фирм, государственных и академических организаций самого высокого уровня. Согласно опубликованному в 2001 году плану Управления по стратегическим инвестициям США, представленному аналитиками ЦРУ, задачей с наивысшем приоритетом (приоритетом А) признан сбор данных из открытых источников, а задачей с приоритетом В – внедрение средств автоматического анализа текстовой информации. Одним из важнейших направлений этой проблематики является классификация различных лингвистических единиц (текстов, слов, словосочетаний, предложений), которая реализуется, практически во всех приложениях лингвистического процессора: информационном поиске, машинном переводе, автоматическом реферировании и др.

На сегодняшний день существует много классификационных систем лингвистических единиц различных уровней языковой системы [1, 2], достаточно много формальных моделей смысловой классификации текстов документов [3, 4], слов [5, 6], реже словосочетаний. Но большинство таких моделей либо не позволяют автоматизировать процедуры

классификации, либо дают высокий уровень шума и низкую точность при их практической реализации. Это связано с тем, что до настоящего времени разработка систем автоматической обработки текстов на естественном языке происходила без использования смыслового анализа или с его минимальным использованием. Для реализации различных этапов лингвистического процессора использовались, в своем большинстве, лексико-грамматические подходы, контекстный анализ, реже синтаксический анализ предложения, и статистические методы и подходы на этапе семантического анализа.

Проведенный анализ показывает, что только использование автоматической классификации, основанной на смысловой близости лингвистических единиц различных уровней языковой иерархической системы, позволяет сократить трудозатраты на поиск нужной информации, повысить полноту и точность выдачи релевантной информации. Один из методов этого перспективного быстро развивающегося направления - метод разбиения пространства лингвистических единиц многоуровневой иерархической языковой системы на группы единиц, имеющих общие элементы смысла, - предложен в настоящей работе.

Постановка проблемы. В работе предлагается метод разбиения пространства лингвистических единиц многоуровневой иерархической языковой системы. Основанием классификации является наличие общих элементов смысла у лингвистических единиц одного уровня языковой системы. Для построения концептуальной схемы понимания смысла вводятся пространство смысловых лингвистических единиц и пространство связных текстов, включающих данные лингвистические единицы. Введение отношений эквивалентности между парами лингвистических единиц и элементами связного текста позволяет факторизовать данные пространств, а использование подходов теории интеллекта позволяет перейти от концептуальной схемы понимания смысла лингвистической единицы к ее реализации в терминах логики предикатов в формальных моделях семантической классификации.

Описание используемого метода. Введем лексикон или множество лингвистических единиц Т. Языковая единица – это элемент системы языка, имеющий различные функции и значения. Совокупности основных языковых единиц образуют «уровни» сложной иерархической языковой системы [7] (например, фонемы – фонемный уровень, морфемы – морфемный уровень). Мы будем рассматривать лингвистические смысловые единицы *t*, начиная от знакового уровня слова к более высоким уровням языковой системы, анализ которых приводит к практическому результату автоматической обработки текстов на естественном языке. Иерархия в системе лингвистических смысловых единиц показана на рис. 1.

Пространство лингвистических смысловых единиц Θ определяется как множество лингвистических единиц лексикона *T*, на котором грамматические правила задают отношения между единицами, выступающими ограничениями для корректных синтаксических структур.

Для определения расстояния β (*t*’, *t*’’) между двумя лингвистическими единицами *t’* и *t’’* будем использовать меру семантической близости *f* (*t’, t’’*)*:*

$$\beta(t', t'') = 1/f(t', t'').$$

Меру семантической близости *f* формально определим соотношением (1) через соответствующие дефиниции глоссариев x_1 и x_2 как мощности множеств, образованных теоретико-множественным пересечением и объединением множеств терминов дефиниций.

$$f(t^{'}, t^{''}) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|} \quad (1)$$

Здесь $x_1 \bigcap x_2$ − общие термины определений, а $x_1 \bigcup x_2$ − все термины определений x_1 и x_2; под термином в данном контексте мы понимаем понятие глоссария в его канонической форме.

Так как для определения семантической близости между понятиями будем использовать несколько словарей, в которых существуют допустимо различные дефиниции одних и тех же лингвистических смысловых единиц, расстояния между двумя лингвистическими единицами удобнее переписать в виде:

$$f(t',t'') = \frac{\sum_{i=1}^{n_1} \frac{\sum_{j=1}^{n_2} f(x_{1i}, x_{2j})}{n_2}}{n_1},$$

где n_1 − количество определений первого термина, взятых из обрабатываемых глоссариев,

n_2 − количество определений второго термина, взятых из обрабатываемых глоссариев,

x_{1i} − i-е определение первого термина, x_{2j} − j-е определение второго термина.

Пусть d − связный текст, включающий лингвистические смысловые единицы $t \in \Theta$. Понятие «связный текст», как объект лингвистической науки допускает множество определений и интерпретаций, которые обусловлены сложностью и многоаспектностью подходов к изучению объекта. Будем понимать под связным текстом законченное информационное и структурное целое, семантически и синтаксически объединяющее смысловой связью последовательность языковых единиц в единый фрагмент. Связный текст представляет собой целостный объект знаковой смысловой единицы верхнего уровня иерархической языковой системы (см. рис. 1). Связный текст может быть представлен высказыванием (реализованным предложением), межфразовым единством (ряд высказываний в едином фрагменте), абзацем, параграфом, главой, разделом, документом и др. [8]

Иерархия отношений элементов связного текста многоуровневой языковой системы наглядно представляется соответствующей теоретико-множественной структурой. Именно, пусть D граф конечного множества связных текстов $\{D_1, D_2, \ldots, D_m\}$, принадлежащих пространству исследуемых связных текстов Ω [135]. Здесь текст $D_i \in \Omega$, $i = 1,\ldots, m$.. При этом текст D_i более высокого уровня иерархии языковой системы можно формально определить через элементы D^j_{i+1} ($D^j_{i+1} \subset D_{i+1}$, j = 1, 2,…..n) связного текста предыдущего уровня иерархии (сверхфразовое единство определяется через фразу, связный текст документа через сверхфразовые единства и т.д.):

$$D_i = \bigcup_{i=1}^{n} D^j_{i+1}, \quad \bigcap_{j=1}^{n} D^j_{i+1} = \varnothing \tag{2}$$

В рассматриваемом пространстве Ω вершина D_i графа D будет родительской для вершин множества $\{D^1_{i+1}, D^2_{i+1}, D^n_{i+1}\}$.

Тогда, расстояние между двумя связными текстами, можно определить как длину пути между соответствующими контекстами $||\alpha(D_i, D_j)||$, определяемую количеством несовпадающих листьев вершин D_i и D_j, а самый короткий путь между двумя элементами связного текста определяется как:

$$||\alpha_{min}(D_i, D_j)|| = \{ \alpha || D_i, D_j || \text{ такое что } \forall ||\alpha`(D_i, D_j)|| \; ||\alpha`(D_i, D_j)|| \geq ||\alpha(D_i, D_j)|| \}$$

Пусть $(t, d) \in (\Theta, \Omega)$ − пара из одной лингвистической смысловой единицы и одного элемента связного текста, где Θ — представляет собой пространство лингвистических

единиц рассматриваемого лексикона *T*, а *Ω* — пространство рассматриваемых связных текстов.

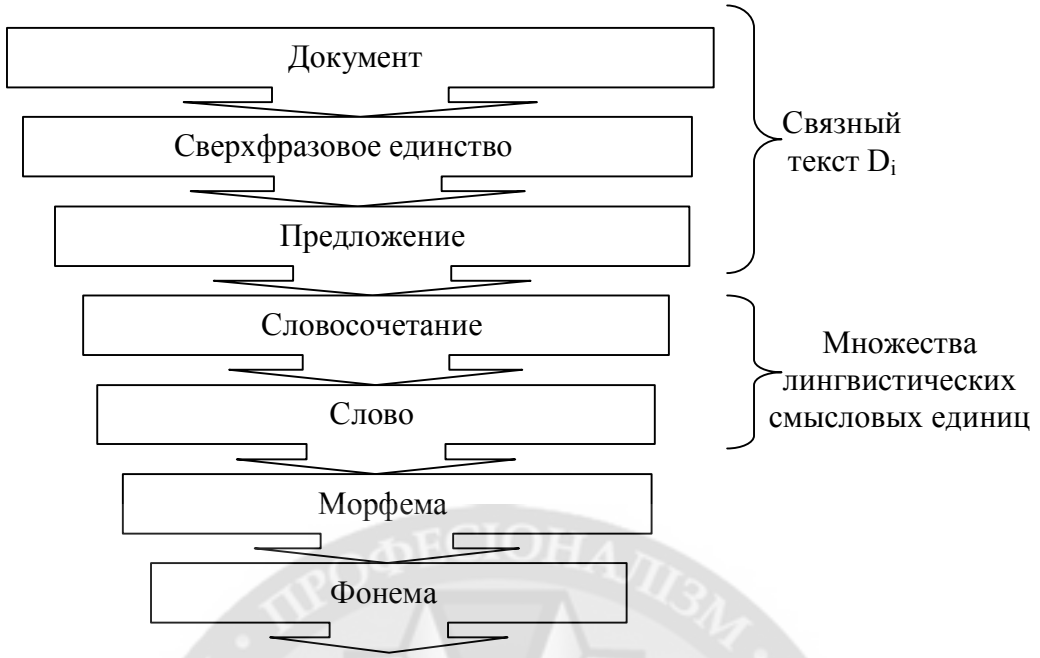

Рис. 1. Структурная схема сложно формализуемой иерархической языковой системы

Если рассмотреть все возможные пары декартового произведения Θ*Ω, то можно построить отображение *F*: (Θ*Ω) → *ϑ*, где ϑ - пространство смысловых полей связных текстов. Схема появления пространства смысловых полей из рассматриваемых связных текстов и привлекаемых лингвистических смысловых единиц представлена на рис. 2.

Таким образом, выбрав лингвистическую смысловую единицу *t* и связный текст *d*, включающий данную лингвистическую единицу, мы определяем смысл элемента связного текста ω через отображение *F*.

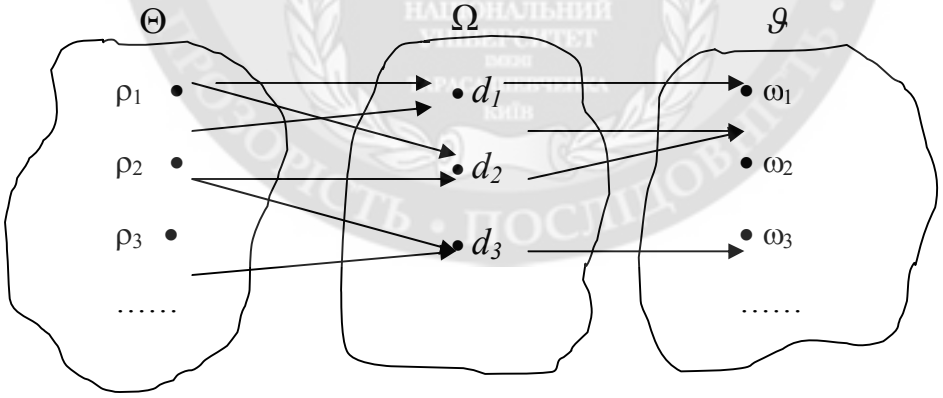

Рис. 2. Схема отображения пространств (Θ*Ω) → *ϑ*

Например:

F (spring, "I made a spring towards a boat")= осуществление человеком прыжка.

F (spring, "He was in the spring of his years") = период жизни человека.

F (spring, "I was in my five and twentieth spring") = период жизни человека.

Введем отображение *G,* такого что (Θ*Ω*)* → *Z*. Здесь *Z* - пространство концептов, выражаемых знаками лингвистических смысловых единиц. Однозначно определив пару *(d, t)* мы приписываем один концепт лингвистической смысловой единице через отображение *G.*

Например,

G (“вечерний наряд может состоять из маленького черного платья”, наряд)= =одежда;

G (“бригаде выдали наряд на работу”, наряд) = распоряжение;

G (“на охрану границы был выслан наряд”, наряд)= подразделение.

Отображение G является однозначным отображением: для каждой пары $(d, t) \in (\Omega, \Theta)$ определяется только один концепт лингвистической смысловой единицы, т.е. в связном тексте лингвистическая единица выражает только одно значение или один концепт.

Пусть $t \in \Theta$ — рассматриваемая лингвистическая единица, а $D_1, D_2, \ldots, D_m$ список элементов анализируемого текста, связанных с данной лингвистической единицей. Тогда имеет место выражение $\forall(t, D_i)\ \exists!\ h \in Z / F^{-1}(t, D_i) = h$.

Будем говорить, что два связных текста контекстуально связаны, и писать $(t`, d`) \sim (t``, d``)$, если только $G(t`, d`) = G(t``, d``)$. Под контекстуальной связностью мы понимаем некую их общность в данном контексте, т.е. отображение некоторого единого семантического поля (некоторого единого смысла или темы) в определенной ситуации языкового окружения или речевого общения.

Будем говорить, что две лингвистические единицы связаны в одном смысле (или в одном своем сигнификативном значении) и писать $(t_i, d_i) \sim (t_j, d_j)$, если только $F(t_i, d_i) = F(t_j, d_j)$.

Например,

F (“application”, “the most Internet applications for the Web are XML-applications”) = “программное обеспечение”;

F (“application”, “application for admission to a university”) = “заявление”;

F (“software”, “using commercial computer-based software”) = “программное обеспечение”;

F (“application”, “the most Internet applications for the Web are XML-applications”) = F (“software”, “using commercial computer-based software”).

Можно показать, что отношение ~ устанавливаемое между лингвистическими смысловыми единицами t и элементами связного текста d, выражает эквивалентность и факторизует пространства лингвистических смысловых единиц Θ и исследуемых связных текстов Ω, разбивая их на классы эквивалентности. Для этого достаточно показать, что отношение ~ является рефлексивным, транзитивным и симметричным [137].

Отношение $(t_i, d_i) \sim (t_j, d_j)$ является рефлексивным отношением. Одна лингвистическая единица в одном своем сигнификативном значении связано само с собой, ибо

$$(t_i, d_i) \sim (t_i, d_i) \leftrightarrow F(t_i, d_i) = F(t_i, d_i).$$

Отношение $(t_i, d_i) \sim (t_j, d_j)$ является симметричным отношением: если одна лингвистическая единица в одном своем сигнификативном значении связана с другой (в одном из ее значений), то вторая лингвистическая единица связана с первой, в вышеназванных значениях.

$$(t_i, d_i) \sim (t_j, d_j) \leftrightarrow F(t_i, d_i) = F(t_j, d_j) \equiv F(t_j, d_j) = F(t_i, d_i) \leftrightarrow (t_j, d_j) \sim (t_i, d_i).$$

Отношение ~ является транзитивным отношением: если одна лингвистическая единица определяет тот же сингификативный смысл, что и вторая, а вторая лингвистической единицей имеет тот же сигнификативный смысл, что и третья, то первая лингвистическая единица в одном из своих сигнификативных значений связана с третьей.

$$(t_i, d_i) \sim (t_j, d_j) \text{ и } (t_j, d_j) \sim (t_k, d_k) \leftrightarrow F(t_i, d_i) = F(t_j, d_j) \text{ и}$$
$$F(t_j, d_j) = F(t_k, d_k) \Rightarrow F(t_i, d_i) = F(t_k, d_k) \leftrightarrow (t_j, d_j) \sim (t_k, d_k).$$

Например,

(“application”, “the most Internet applications for the Web are XML-applications”) ~ (“software”, “using commercial computer-based software”) и *(“software”, “using commercial computer-based software”) ~ (“program”, “everything done on a computer is done by using a program”) ↔ F(“application”, “the most Internet applications for the Web are XML-*

applications")= F ("software", "using commercial computer-based software")= F("program", "everything done on a computer is done by using a program") ="программное обеспечение".

Данное отношение эквивалентности позволяет организовать различные пары лингвистических единиц и связных текстов, включающих данные единицы, *(t, d)*, в классы эквивалентности, определяющими один и тот же сигнификативный смысл, тем самым, факторизовав пространство концептов, выражаемых знаками лингвистических смысловых единиц, $(\Omega, \Theta) \to Z$.

$$[(t_j,d_j)] = \{(t_i,d_i) \in(\Omega,\Theta) /(t_i, d_i) \sim (t_j,d_j) \}\equiv\{(t_j,d_j) \in(\Omega x\Theta)/ F (t_i, d_i) = F (t_j, d_j)\}.$$

Отношение эквивалентности ~ делает F однозначным отображением, в котором два класса имеют одинаковое сингификативное значение, если они являются одним и тем же классом, что позволяет нам выбрать одну репрезентативную лингвистическую единицу, представляющую подходящее значение из каждого класса эквивалентности.

Введение контекстно-знакового предиката. Для реализации данного метода используем подходы теории интеллекта [9] и разработаем модель интеллектуально понимания смысла. На декартовом произведении элементов множеств $T * D$ вводим контекстно-знаковый предикат $L(t_i, d_j)$, задающий отношения между лингвистическими единицами лексикона и контекстом (или связным текстом). Если $L (t_i, d_j) = 1$, то это значит, что лингвистическая единица t_i из множества T однозначно соответствует обрабатываемому связному тексту $d_j \in D$. Если $L (t_i,d_j) = 0$, то t_i не соответствует связному тексту d_j.

Предикат L должен удовлетворять постулату существования: предикат $L(t_i, d_j)$ реально существует в том и только в том случае, если при повторном предъявлении любой пары (t_i, d_j) из множества $T * D$ всегда будет получен тот же ответ, что и в первый раз.

Таким образом, контекстно-знаковый предикат $L(t_i, d_j)$ для каждой пары t_i и d_j объективно отображает отношение включение знака лингвистической смысловой единицы в элемент связного текста.

Отношение эквивалентности ~, факторизующее пространство лингвистических смысловых единиц Θ, разбивая его на классы эквивалентности, однозначно определяется контекстно-знаковым предикатом $L(t_i, d_j)$ и позволяет ввести предикаты эквивалентности:

- предикат интегральных семантических признаков лингвистических единиц G_t, заданный на декартовом квадрате $T * T$:

$$G_t (t', t'') = \forall d \in D (L(t', d) \sim L(t'', d)).$$

Предикат $G_t (t', t'')$ можно использовать для объективного определения общих интегральных семантических признаков лингвистических единиц. Действительно, если $G_t (t', t'') =1$, то $L(t', d) = L(t'', d)$ и если $G_t (t', t'') =0$, то $L(t', d) \neq L(t'', d)$ для любого связного текста $d \in D$, т.е. две лингвистические единицы в том же контексте либо имеют один или более общие интегральные семантические признаки (один или более), либо не имеют таких признаков.

Например, если G_t *("application","software")=1*, то *L (application, "using commercial computer-based application") = L ("software","using commercial computer-based software")*.

- и предикат семантического поля текста G_d, заданный на декартовом квадрате $D * D$:

$$G_d(d', d'')= \forall t \in T (L(d', t) \sim L(d'', t)).$$

Предикат $G_d(d', d'')$ можно использовать для объективного определения принадлежности связных текстов к единому семантическому полю (некоторому единого смысла или темы). Действительно, если $G_d(d', d'')=1$, то при любом понятии, выражаемом лингвистической смысловой единицей из множества $t \in T$, связные тесты d' и d'' будут относиться к единому семантическом полю, обладая некой синонимизацией, т.е. соответствием одной естественно-тематической группе. И, наоборот, если $G_d(d', d'')=0$, то два связанных текста с одной и той же лингвистической смысловой единицей будут соответствовать различным семантическим полям.

Например, если в качестве связного текста используются фразы *d'="вечерний наряд может состоять из маленького черного платья"* и *d''="на охрану границы был выслан наряд"*, включающие понятие *="наряд"*, то $G_d(d', d'')=0$, и соответственно d' и d''относятся к различным семантическим полям.

Предикат G_t определяет разбиение Ψ_t множества T на слои лингвистических единиц. Все лингвистические единицы, принадлежащие одному слою разбиения, относятся к синонимичным концептам, а любые лингвистические смысловые единицы, взятые из разных слоев разбиения Ψ_t, относятся концептам, не имеющим общих элементов смысла.

Предикат G_d определяет разбиение Ψ_d множества D на слои связных текстов. Все связные тексты, принадлежащие одному слою разбиения, относятся к одной естественно-тематической группе. А любые два связных текста, взятые из разных слоев разбиения Ψ_d относятся к различным естественно-тематическим группам тестов.

Выводы. Таким образом, предлагаемый метод позволяет произвести разделение пространства исследуемых лингвистических единиц на группы, имеющие некоторые общие элементы смысла, т.е. факторизовать пространство лингвистических единиц различных уровней иерархической языковой системы по семантическим парадигматическим признакам, выделяя, например, семантические эквиваленты слов и словосочетаний. Метод осуществляет также факторизацию пространства связных текстов для отнесения единицы связного текста к различным семантическим полям или подклассам различных тематик.

Введение и обоснование контекстно-знакового предиката позволяет формально представить отношения между смысловой лингвистической единицей, выражаемой знаком, и включающим ее связным текстом. Результатом использования предиката интегральных семантических признаков лингвистических единиц и предиката семантического поля текста является разбиение множества лингвистических смысловых единиц и множества связных текстов на слои эквивалентности, различаемые по семантическим признакам.

ЛИТЕРАТУРА:

1. Мельчук И.А. Курс общей морфологии. Том III. Часть третья: морфологические средства; Часть четвёртая: морфологические синтактики. Москва: Языки русской культуры – Вена : Wiener Slawistischer Almanach. 2000. − 367 с.

2. Глагольные лексемы со значением качественного и качественно-оценочного признака :на материале современного немецкого языка. Дисс. Фадеева Л. В. на соискание ученой степени кандидата филологически наук. 2007, Омск.

3. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа / Андреев А. М., Березкие Д. В., Морозов В. В., Симаков К. В. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Пятой всероссийской научной конференции (RCDL'2003). – СПб.: НИИ Химии СПбГУ, 2003. – С.140-149.

4. Сбойчаков К.О. Автоматизированная система смысловой обработки текстов при создании электронных фондов библиотеки. Дисс. на соискание ученой степени кандидата технических наук. 2003, Москва. − 178 с.

5. Апресян Ю.Д. Избранные труды. Том I. Лексическая семантика (синонимические средства языка). 2-изд., испр. и доп. – М.: Языки русской культуры, 1995. – 464 с.

6. Кретов А.А., Рафаева А.В. Программа семантической классификации ПроСеКа: теоретические и прикладные аспекты // 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.230-235.

7. Ахманова О.С. Единицы языка // Словарь лингвистических терминов. – Изд. 4-е, стереотипное. – М.: КомКнига, 2007. – 576 с.

8. . Ерофеева Е.В., Кудлаева А.Н. К вопросу о соотношении понятий «текст» и «дискурс» // Проблемы социо- и психолингвистики: Сб. ст. Вып. 3. Пермь, 2003.

9. Шабанов-Кушнаренко Ю.П., Шаронова Н.В. Компараторная идентификация лингвистических объектов: Монография. – К.: ИСДО, 1993. – 116 с.

Рецензент: д.т.н., проф. Замаруєва І.В.