Теоретические основы, методы и алгоритмы формирования знаний о синонимии для задач анализа и сжатия текстовой информации

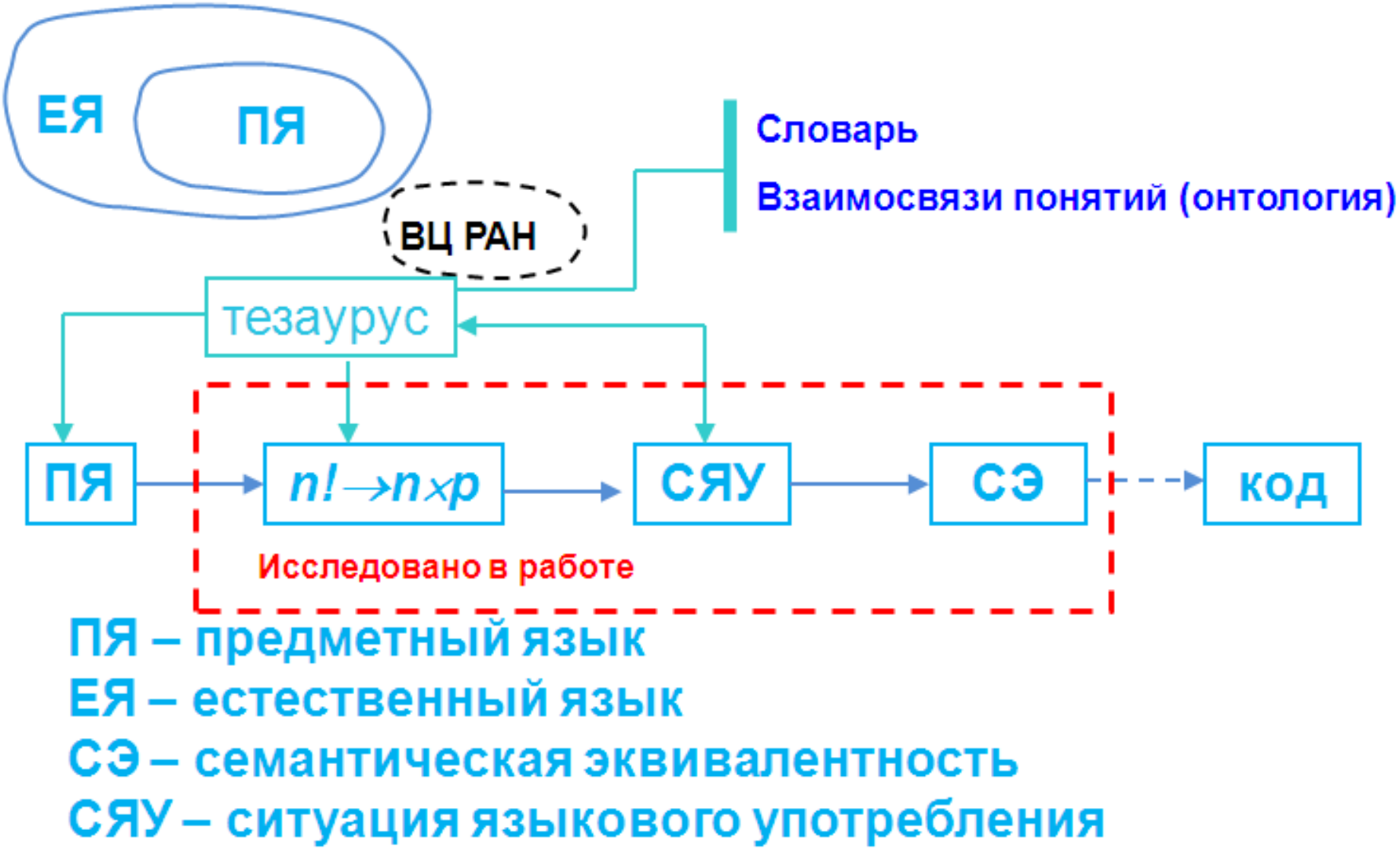
Михайлов Дмитрий Владимирович

Диссертация на соискание учёной степени
доктора физико-математических наук

05. 13. 17 – теоретические основы информатики

Научный консультант: д. т. н., профессор Емельянов Г.М.

Великий Новгород, 2012 г.



ПЯ – предметный язык
ЕЯ – естественный язык
СЭ – семантическая эквивалентность
СЯУ – ситуация языкового употребления

Объект, предмет и цель исследования

Объект исследования

Программные средства распознавания, анализа и сжатия текста на естественном языке (ЕЯ).

Предмет исследования

Методы и алгоритмы формирования знаний о синонимии.

Основная цель диссертационной работы

Разработка теоретических основ организации структуры и использования знаний о синонимии для совокупности задач оценки семантической схожести текстов предметно-ограниченного ЕЯ, автоматизации пополнения и компрессии баз языковых и предметных знаний.

Задачи исследования

Основные задачи исследования

- Анализ известных методов формализации семантики конструкций естественного языка и определение функциональных требований к механизму сравнения смыслов.
- Разработка и исследование методов анализа семантической эквивалентности (СЭ) на уровне варьирования абстрактной лексикой.
- Разработка методов автоматизированного формирования и кластеризации знаний о семантике конструкций предметно-ограниченного ЕЯ с учётом взаимосвязи языковых уровней.
- Исследование и алгоритмизация механизма использования морфологии и синтаксиса ЕЯ для задач кластеризации, разделения и сжатия баз предметных и языковых знаний
- Разработка и исследования методов численной оценки семантической схожести текстов предметно-ограниченного ЕЯ.
- Разработка архитектуры программной системы контроля знаний, реализующей предложенные принципы, методы и алгоритмы.

Основные задачи и результаты первой главы

Основные задачи:

1. Анализ известных методов формализации семантики языковых конструкций.
2. Формулировка общих требований к механизму установления семантической эквивалентности текстов предметно-ограниченного подмножества естественного языка.

Важнейшие результаты (п. 6 паспорта специальности 05. 13. 17):

1. Концепция ситуации языкового употребления (СЯУ) как единицы формализованного описания семантики естественного языка.
2. Комплексная методика формирования и экспериментальной оценки знаний в виде классов смысловой эквивалентности текстов на основе ситуаций употребления предметно-ограниченного подмножества естественного языка.

Определение 1.4

Ситуация Языкового Употребления (СЯУ) — описание нового социального опыта (содержания совместных действий) средствами заданного ЕЯ.

Фиксируемый СЯУ S языковой контекст представляется тройкой:

$$S = (O, R, Ts)\,, \tag{1.1}$$

где O — множество символов, отождествляемых с некоторыми понятиями;
Ts — множество форм описания S в некоторой знаковой системе;
$R \subset O^n$, где $n \in 1, \ldots, |O|$.

Пусть $Synt$ — сюръективная функция, определяемая синтаксисом языка;

Тогда для $\forall Ts_i \in Ts \;\; \exists Tr_i \colon Ts_i = Synt\,(Tr_i)$, Tr_i — помеченное дерево.

При этом если $O = M \cup V$, $M \cap V \neq \emptyset$, то для $\forall o_j \in M$ найдётся $o_k \in V$ такое, что понятию o_j соответствует дочерний узел с пометкой w_j, а понятию o_k — родительский узел с пометкой w_k в дереве Tr_i.

Задача 1.1

Дано:

- Множество ЕЯ-текстов G.

Требуется: по результатам разбора каждого $g_i \in G$ выявить:

- Множества $V(g_i)$ и $M(g_i)$.
- Тернарное отношение $I \subseteq G \times M \times V$: $M = \bigcup_i M(g_i)$, $V = \bigcup_i V(g_i)$.

На основе I необходимо сформировать множество R и выделить группы текстов по сходству встречаемости понятий в одних и тех же $r_j \in R$.

Пусть $A \subseteq G$, $B \subseteq M \times V$ и существует пара отображений:

$$A' = \left\{(m, v) : m \in M, v \in V \;\middle|\; \forall g \in A \colon m(g) = v\right\},$$
$$B' = \left\{g \in G \;\middle|\; \forall (m, v) \in B \colon m(g) = v\right\}.$$

Определение 1.11

Пара (A, B), где $A' = B$ и $B' = A$, есть формальное понятие (ФП) с объёмом A и содержанием B.

При этом каждому классу СЭ соответствует некоторый класс формальных понятий в решётке, а задача накопления знаний о синонимии сводится к совокупности следующих подзадач:

- формирование прецедентов синонимии для уровня абстрактной лексики.
- кластеризация отношений из множества R в составе тройки (1.1).
- численная оценка схожести СЯУ.

Основные задачи и результаты второй главы

Основные задачи:

1. Описание на функциональном уровне задачи увеличения полноты описания смысла на уровне глубинного синтаксиса.
2. Разработка теоретических основ выделения сверхфразовых единств на указанном языковом уровне.

Важнейший результаты (пп. 4 и 10 паспорта специльности 05. 13. 17):

1. Теоретическое обоснование разрешимости задачи построения системы целевых выводов в грамматике деревьев на основе её информационно-логической модели.
2. Метод выделения сверхфразовых единств в текстах на уровне глубинного синтаксиса как основа сжатия смысловой информации для данного языкового уровня.

Определение

Лексико-синтаксическая грамматика деревьев (далее — грамматика деревьев, Δ-грамматика) задаётся посредством четвёрки вида

$$\Gamma = (W_R, V_R, \phi, \Pi),$$

где V_R — конечное множество пометок на ветвях, $V_R = \{a_1, a_2, \ldots, a_k\}$;

W_R — конечное множество пометок на узлах деревьев;

ϕ — матрица ограничений на характер размещения пометок из V_R, для $\forall\, i = 1, \ldots, k$ из любого узла дерева выходит не более $\phi(a_i) = n_i$ ветвей с пометкой a_i;

Π — конечное множество правил преобразований деревьев, причём для $\forall\, rule_j \in \Pi$ задаётся множество Rap условий его применимости.

Замечание

Содержательно $\forall\, rap_i \in Rap$ выступает в роли прецедента, с которым отождествляется класс СЭ на уровне абстрактной лексики.

Определение 2.1

Лексической синонимической конструкцией (ЛСК) далее именуется комплекс лексических единиц $wr_k \in W_R$ и связей $vr_j \in V_R$ между ними, замена которого описывается некоторым $rule_i \in \Pi$.

Каждой ЛСК соответствует своё ключевое слово C_0, при этом в общем случае произвольная wr_k в составе ЛСК есть значение некоторой лексической функции (ЛФ) от C_0.

Представим вход правила $rule_j \in \Pi$ как описание заменяемого поддерева.

Тогда для любого $rule_j \in \Pi$ результат анализа применимости к заданному дереву фиксируется списком пар:

$$\{(rule_j, C_0(j) : j = 1, \ldots, |\Pi|)\}, \tag{2.1}$$

причём в работе $\forall rule_j \in \Pi$ выделяются два состояния: для заменяемого дерева Tio_1 и для заменяющего дерева Tio_2, $Tio_k = \langle Wio_k, Vio_k \rangle$, где Wio_k и Vio — множества узлов и ветвей, соответственно.

Обозначим $\vee_{l=1}^{m} rap_l$ как r_{12}. Тогда применение $rule_j \in \Pi$ есть переход:

$$rule_j\left(r_{12}\right) : Tio_1 \xrightarrow{rule_j(r_{12})} Tio_2. \tag{2.7}$$

Утверждение

В общем случае определяемый правилом $rule_j \in \Pi$ переход из Tio_1 в Tio_2 допустим, если $\exists rap_l \in Rap$: $\vee_{l=1}^{m} rap_l = \text{true}$, где $m = |Rap|$.

Отдельному правилу соответствует элементарная сеть Петри вида

$$N = \{P, T, F, H, M_0\}, \tag{2.8}$$

где P — множество позиций, $P = \{p_1, p_2\}$, $p_1 \Leftrightarrow Tio_1$, $p_2 \Leftrightarrow Tio_2$;

T — множество возможных переходов, $T = \{t\}$, $t = rule_j\left(r_{12}\right) : p_1 \xrightarrow{t} p_2$;

F и H — отображения, $F\colon P \times T \to \{0,1\}$, $H\colon T \times P \to \{0,1\}$,
для сети (2.8) $F\left(p_1, t\right) = 1$, $F\left(p_2, t\right) = 0$, $H\left(t, p_1\right) = 0$, $H\left(t, p_2\right) = 1$;

M_0 — вектор начальной маркировки, $M_0 = (1,0)$,
второй допустимой маркировке соответствует вектор $M = (0,1)$.

Рассмотрим $\Pi_R \subseteq \Pi$: для $\forall\, rule_1 \in \Pi_R \;\; \exists\, rule_2 \in \Pi_R,\; rule_2 \neq rule_1$: либо вход $rule_2$ является выходом $rule_1$, либо вход $rule_1$ есть выход $rule_2$.

Пусть N_i — сеть Петри, построенная из примитивов, каждый из которых моделирует работу некоторого $rule_j \in \Pi_R$.

Тогда последовательность применения правил $rule_j \in \Pi_R$ соответствует последовательности $\tau = (t_{1i}, t_{2i}, \ldots, t_{ki})$ срабатываний переходов:

$$Tio_1 \xrightarrow{rule_1(r_{12})} Tio_2 \xrightarrow{rule_2(r_{23})} Tio_3 \to \cdots \to Tio_k \xrightarrow{rule_k\left(r_{k,\,k+1}\right)} Tio_{k+1}. \quad (2.9)$$

При этом происходит последовательная смена разметок:

$$M_{0i} \xrightarrow{t_{1i}} M_{1i} \xrightarrow{t_{2i}} M_{2i} \to \cdots \to M_{k-1,i} \xrightarrow{t_{ki}} M_{ki}, \quad (2.10)$$

где $t_{1i} \Leftrightarrow rule_1\,(r_{12}),\;\; t_{2i} \Leftrightarrow rule_2\,(r_{23}),\;\; \ldots\;,\;\; t_{ki} \Leftrightarrow rule_k\,(r_{k,\,k+1}),$

$M_{0i} \Leftrightarrow Tio_1,\;\; M_{1i} \Leftrightarrow Tio_2,\;\; \ldots\;,\;\; M_{k-1,i} \Leftrightarrow Tio_k,\;\; M_{ki} \Leftrightarrow Tio_{k+1}.$

Замечание

Множество достижимости сети N_i зависит от задания M_{0i}.

Пусть T_i — множество переходов сети N_i, рассматриваемое как алфавит.
Тогда приведение Tio_1 и Tio_{k+1} к виду с одинаковой ЛСК включает:

- определение достижимости разметки M_{ki} из начальной разметки M_{0i}.
  Данная задача есть поиск слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_{ki}$,
  где T_i^* — множество всех слов в алфавите T_i языка $L(N_i)$;
- задачу обратимости слова τ: если $\tau \in T_i^*$, то существует ли
  слово $\tau' = (t'_{ki}, t'_{k-1,\,i}, \ldots, t'_{2i}, t'_{1i})$:

$$M_{0i} \xleftarrow{t'_{1i}} M_{1i} \xleftarrow{t'_{2i}} M_{2i} \leftarrow \cdots \leftarrow M_{k-1,\,i} \xleftarrow{t'_{ki}} M_{ki}, \qquad (2.11)$$

  где $M_{0i} \Leftrightarrow Tio_1$, $M_{1i} \Leftrightarrow Tio_2$, $\ldots$, $M_{k-1,\,i} \Leftrightarrow Tio_k$, $M_{ki} \Leftrightarrow Tio_{k+1}$;
- задачу определения оптимального слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_{ki}$.
  Суть: если существуют $\tau_1, \tau_2, \ldots, \tau_l$, описывающие смену разметок

$$M_{0i} \xrightarrow{\tau_1} M_{ki}, \; M_{0i} \xrightarrow{\tau_2} M_{ki}, \; \ldots, \; M_{0i} \xrightarrow{\tau_l} M_{ki},$$

  то оптимальное слово есть обратимое слово минимальной длины.

Лемма 2.2

Проблема достижимости заданной разметки M_{ki} из начальной M_{0i} в сети N_i разрешима.

Теорема 2.1

Сеть N_i безопасна в течение всего времени функционирования моделируемой системы правил Δ-грамматики.

Теорема 2.3

Проблема определения обратимости слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_{ki}$ языка $L(N_i)$ является разрешимой (T_i^* — множество слов в алфавите T_i).

Теорема 2.4

Проблема поиска оптимального слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_{ki}$ разрешима.

Основные задачи и результаты третьей главы

Основные задачи:

1. Формализация понятия прецедента ситуации синонимии на уровне абстрактной лексики.
2. Автоматизации накопления и экспериментальной оценки знаний об условиях применимости синонимических преобразований деревьев глубинного синтаксиса, рассмотренных во второй главе.
3. Разработка комплексной методики формирования и кластеризации семантических отношений в рамках концепции ситуации языкового употребления.

Важнейший результаты (пп. 5 и 6 паспорта специльности 05. 13. 17):

1. Разработан принцип формирования и кластеризации синтаксических отношений выделением синтагматических зависимостей в рамках ситуации языкового употребления.
2. Программная реализация указанного принципа (имеется свидетельство об официальной регистрации программы для ЭВМ).

Пусть Ts — множество СЭ-фраз, задающих некоторую СЯУ согласно (1.1).

При рассмотрении $\forall Ts_i \in Ts$ в её составе следует выделить:

- неизменную часть, общую для всех $Ts_i \in Ts$;
- изменяемую (флективную) часть.

При этом для $\forall Ts_i \in Ts$ справедливо то, что $Ts_i = \bigodot\limits_j W_{ij}$,

где $\odot$ есть операция конкатенации символьных последовательностей, а

$$W_{ij} = Wc_{ij} \odot Wf_{ij},$$

где Wc_{ij} — неизменная часть (основа);

Wf_{ij} — флективная часть.

На основе попарного сравнения W_{ij} различных Ts_i требуется найти:

- Wc_{ij} и Wf_{ij} каждого W_{ij} при $|Wc_{ij}| \to \max$;
- синтаксическое отношение R_q, определящее допустимость сочетания слов с буквенным составом флексий Wf_{ij} и Wf_{ik}, $k \neq j$.

Пусть J — индексное множество для неизменных частей всех слов, употребленных во всех $Ts_i \in Ts$.

Определение 3.2

Упорядоченная совокупность индексов $j \in J$ неизменных частей слов, входящих в Ts_i, будет *моделью линейной структуры* этой фразы, $Ls\,(Ts_i)$.

Пусть $h\,(j, Ls\,(Ts_i))$ — позиция индекса j в модели $Ls\,(Ts_i)$.

Тогда множество связей для $Ls\,(Ts_i)$

$$D : Ts_i \to \left\{ \Big(h\big(j, Ls\,(Ts_i)\big),\ h\big(k, Ls\,(Ts_i)\big) \Big) \ :\ j \neq k \right\}.$$

Определение 3.3

Связь $d_{qi} = \Big(h\big(j, Ls\,(Ts_i)\big),\ h\big(k, Ls\,(Ts_i)\big) \Big)$ считается допустимой для модели $Ls\,(Ts_i)$, если $\exists\ \{Ts_l, Ts_m\} \subset Ts,\ l \neq m,$ причём и $Ls\,(Ts_l)$, и $Ls\,(Ts_m)$ имеют подпоследовательностью либо $\{j, k\}$, либо $\{k, j\}$.

При этом пара $(j,\ k)$ содержательно соответствует одной синтагме.

Пусть для $\forall Ts_i \in Ts,\ i = 1, \ldots |Ts|$, все $d_{qi} \in D\left(Ts_i\right)$ удовлетворяют *определению 3.3*.

Определение 3.4

Будем считать, что модель $Ls\left(Ts_i\right)$ проективна относительно множества синтаксических связей в Ts, если

$$\sum_{q=1}^{|D(Ts_i)|} \Delta_{qi} \leq \left|Ls(Ts_i)\right|, \text{ где}$$

$$\Delta_{qi} = \left|h\big(j, Ls\left(Ts_i\right)\big) - h\big(k, Ls\left(Ts_i\right)\big)\right|.$$

Замечание

Формирование множества связей, отвечающих *определению 3.3*, в общем случае предполагает порядка $mn^2 + m^2n$ операций типа сравнения пары слов по буквенному составу, где $m = \left|Ts\right|$, $n = \max\limits_{i=1,\ldots,|Ts|}\left(\left|Ts_i\right|\right)$.

Отбор СЭ-фраз, модели линейных структур которых удовлетворяют *определению 3.4*, дополнительно требует порядка $2nm\left(\frac{m(n-1)}{2}\right)^2$ таких операций (доказательство приводится в тексте диссертации, с. 136–138).

Пусть $\bigcup_i D(Ts_i)$ есть множество связей, допустимых в рамках Ts.

Определение

Множество пар (j, k), сгруппированных по некоторому общему индексу k, есть элемент множества V_J вершин графа синтагм (V_J, I_J). При этом множества E_1 и E_2, входящие в V_J, будут соединены ребром из I_J, если $\exists\, \{j, k, m\} \subset J$: $(j, k) \in E_1$, $(k, m) \in E_2$ и $j \neq m$.

Анализом (V_J, I_J) строится дерево синтаксических связей (V_{JT}, I_{JT}). Формально

$$V_{JT} = J, \quad I_{JT} = \Big\{(j, k) : \exists\, E \in V_{JT},\, (j, k) \in E\Big\} \tag{3.11}$$

При этом $k \in V_{JT}$ соответствует корню дерева (3.11), если $\exists\, E_1 \in V_J$, в котором пары индексов сгруппированы по k, $|E_1| > 1$, а k не содержится ни в одной паре индексов для $\forall\, E_2 \in V_J$: $E_1 \neq E_2$.

Замечание

Число дочерних узлов у корня дерева (3.11) полагается не менее двух, поскольку содержательный интерес для выделения отношений в рамках СЯУ представляют ситуации с двумя и более участниками.

Пусть $T_i^{\odot} = \left\{w_{ij} : w_{ij} = \odot\left(W_{ij}\right)\right\}$. Положим также, что $\exists\, Tp_i \subset Ts_i$, определяющее последовательность

$$P_i^{\odot} = \left\{u_k : u_k = \odot\left(Wp_k\right), \bigcup_k Wp_k = Tp_i\right\},$$

где $Wp_k \in Ts_i$ — последовательность символов слова, для которого не выделены неизменная и флективная часть

Теорема 3.1

Последовательность $P_i^{\odot}$ содержит предикатное слово, если

$$\exists\, \{j, 0, k\} \subset Ls\left(Ts_i\right) : \{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{\odot},$$

где $\{u_1, \ldots, u_p\} = P_i^{\odot}$, а $p = \left|P_i^{\odot}\right|$.

Теорема 3.2

Слово $u_k \in P_i^{\odot}$ принадлежит расщепленному предикатному значению (РПЗ), если $\exists\, Ts_j \in Ts$: $Ls\left(Ts_j\right) \neq Ls\left(Ts_i\right)$, а $u_k \in P_j^{\odot}$, причём $P_j^{\odot}$ также отвечает условию теоремы 3.1. При этом $\nexists\, Ts_k \in Ts$, где $P_k^{\odot} \subset P_i^{\odot}$ и отвечает теореме 3.1, а $Ls\left(Ts_k\right) \neq Ls\left(Ts_j\right)$ и $Ls\left(Ts_k\right) \neq Ls\left(Ts_i\right)$.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих *теореме 3.2*, а $Ts' \subset Ts$, при этом $Ts' = \left\{Ts_i \colon \left|P_i^{\odot\prime}\right| \to \max\right\}$.

Для $u_k \in \bigcup_i P_i^{\odot\prime}$, $Ts_i \in Ts'$, неизменная и флективная части формируются сравнением последовательности Wp_k для всех $u_j \in \bigcup_l P_l^{\odot}$: $Ts_l \in (Ts \setminus Ts')$, а $P_l^{\odot}$ отвечает условию *теоремы 3.1*.

При этом необходимо, чтобы $2\,|Wc_k| > |Wf_k| + |Wf_j|$, где $Wp_k = Wc_k \odot Wf_k$, а $Wp_j = Wc_k \odot Wf_j$.

С учётом $P_i^{\odot\prime}$ дерево (3.11) преобразуется следующим образом:

- корень изменяется с $k = 0$ на значение k для $u_k \in P_i^{\odot\prime}$ с максимальной встречаемостью в разных ЕЯ-фразах из Ts;
- правое поддерево перевешивается на узел j для слова $u_j \in P_i^{\odot\prime}$ наименьшей встречаемости;
- в паре $\{u_l, u_m\} \subset P_i^{\odot\prime}$ дочерний узел у слова меньшей встречаемости.

Определение 3.5

Дерево вида (3.11), преобразованное согласно указанным правилам, назовём далее расширенным деревом (3.11). Фактически оно есть дерево-прецедент для множества деревьев $\{Tr_i \colon Ts_i = Synt\,(Tr_i)\,, Ts_i \in Ts\}$.

Основные задачи, решаемые в четвёртой главе:

1. Разработка принципов выделения и классификации частичных смысловых эквивалентностей по результатам разбора текстов внешней программой синтаксического анализа.
2. Изучение алгоритмических аспектов механизма использования морфологии и синтаксиса естественного языка для формирования классов смысловой эквивалентности текстов.

Важнейший результаты (п. 6 паспорта специльности 05. 13. 17):

1. Алгоритмы формирования объектно-признакового описания ситуаций языкового употребления согласно постановке *Задачи 1.1* с примененем внешней программы синтаксического анализа.
2. Комплексная методика формирования и экспериментальной оценки знаний на основе синтаксического контекста существительного.

Основные задачи и результаты пятой главы

Основные задачи:

1. Развитие предложенного в третьей главе принципа выделения и кластеризации синтаккcических отношений на случаи наличия синонимов в составе семантически эквивалентных фраз из определяющих ситуацию языкового употребления.
2. Разработка метода численной оценки схожести ситуаций употребления предметно-ограниченного естественного языка.

Важнейший результаты (п. 5 паспорта специльности 05. 13. 17):

1. Теоретико-решёточная концепция ситуации языкового употребления как информационной единицы тезауруса предметной области.
2. Метод численной оценки семантической схожести текстов предметно-ограниченного естественного языка относительно ситуаций его употребления.

Пусть LS — множество моделей линейных структур ЕЯ-фраз из Ts на J.

Теорема 5.1

Пара индексов $\{j_1, j_2\} \subset J$ соответствует словам-синонимам, если $\exists \{Ls(Ts_1), Ls(Ts_2)\} \subseteq Ls$:

$$Ls(Ts_1) = J_1 \odot \{j_1\} \odot J_2 \text{ и } Ls(Ts_2) = J_1 \odot \{j_2\} \odot J_2,$$

где $J_1 \subset J$, $J_2 \subset J$, а $\odot$ есть операция типа конкатенации над множеством J.

Пусть PJ — множество пар, отвечающих *теореме 5.1*, а множество LS' формируется из LS заменой индексов, вошедших в пары из PJ, на некоторые $j \in (\mathbb{N} \backslash J)$. Обозначим множество заменяемых индексов как JP, а множество индексов замены — как JP', $JP' \cap JP = \varnothing$.

Утверждение 5.1

Индексы с максимальной встречаемостью в LS' соответствуют словам-существительным, обозначающим участников ситуации (1.1).

Обозначим множество индексов, отвечающих *утверждению 5.1*, как JN. Пусть $Ls_1(Ts_i) \in LS'$, а $Ls_2(Ts_i)$ — модель линейной структуры Ts_i, но относительно JN. Обозначим множество моделей второго вида как LJN.

Положим, $\exists\, LS'_j \subset LS'$: для $\forall\, Ls_1(Ts_i) \in LS'_j$ все $Ls_2(Ts_i)$ одинаковы и соответствуют некоторой $Ls_2(Ts_j) \in LJN$, $Ts_j \in Ts$.

Утверждение 5.2

Индексы $j \notin JN$ с максимальной встречаемостью в различных $Ls_1(Ts_i) \in LS'_j$ могут соответствовать наречиям, прилагательным, либо опорным существительным в составе генитивных конструкций.

Множество таких индексов обозначим далее как JA.

Замечание

Местоположение индекса в расширенном дереве (3.11) и выделение флексий для слов с индексами из $\big((J \backslash JP) \cup JP'\big) \backslash (JN \cup JA) \cup \{0\}$ производится по аналогии с выделением указанной информации у слов, отвечающих *теореме 3.2*. При этом вместо индексов с ненулевым значением рассматриваются индексы из $JN \cup JA$.

Представим СЯУ посредством формального контекста (ФК):

$$Ks = (Gs, Ms, Is), \tag{5.1}$$

где $\forall g \in Gs$ — основа слова, синтаксически подчинённого другому слову из некоторой $Ts_i \in Ts$ в составе тройки (1.1).

Множество признаков Ms включает подмножества, обозначаемые далее посредством соответствующего нижнего индекса и содержащие:

- указания на основу синтаксически главного слова (индекс 1);
- указания на флексию главного слова (индекс 2);
- связи «основа–флексия» для главного слова (индекс 3);
- сочетания флексий зависимого и главного слова (индекс 4). При этом после флексии главного слова через двоеточие указывается предлог (если такой имеется) для связи главного слова с зависимым;
- указания на флексию зависимого слова (индекс 5).

Посредством $Is \subseteq Gs \times Ms$ выделяются классы отношений из R в (1.1) по сходству основы главного, флексии зависимого слова, лексической и флективной сочетаемости.

Рассмотрим представление тезауруса формальным контекстом

$$Kth = (Gth, Mth, Ith)\,, \tag{5.2}$$

где Gth состоит из символьных пометок отдельных СЯУ.

Mth содержит признаки формальных контекстов всех $gth \in Gth$.

Кроме того, в составе Mth выделяются подмножества:

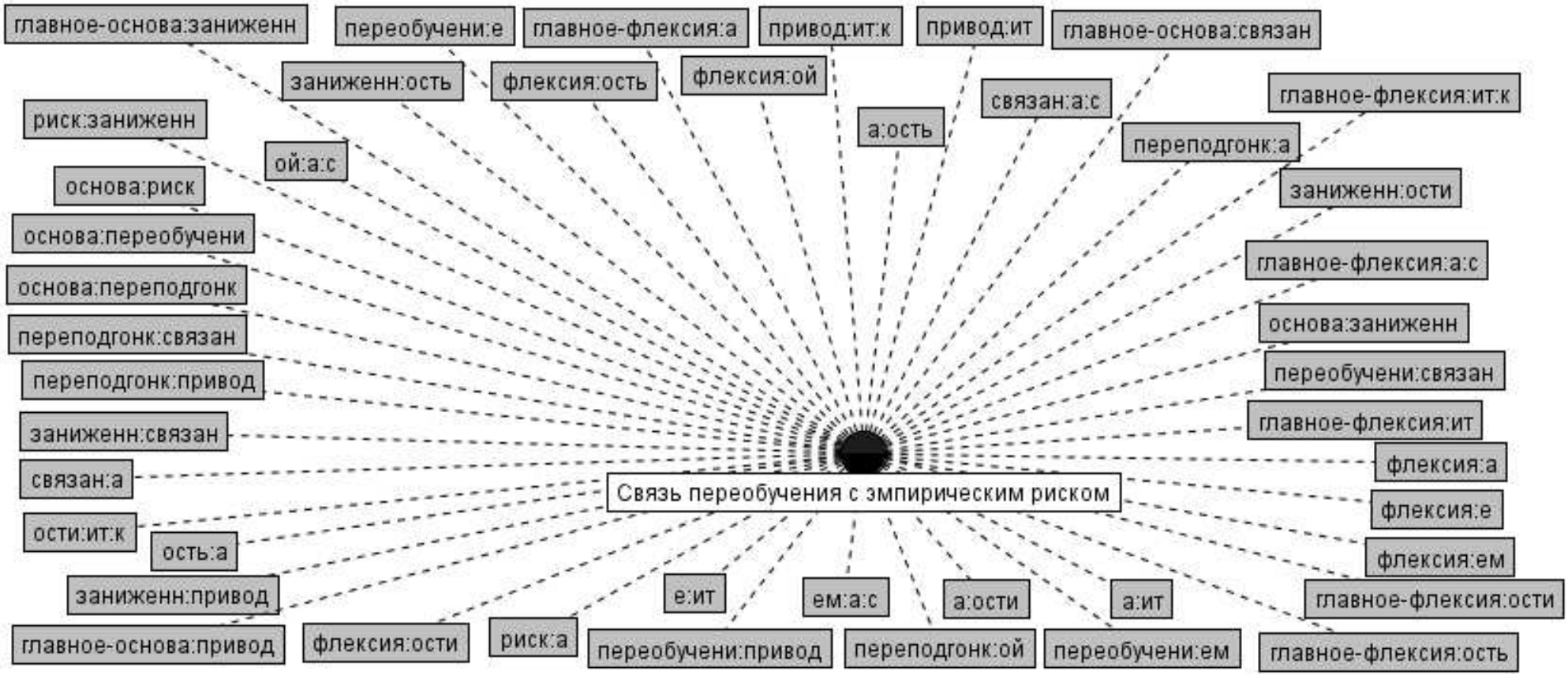
- M_6 — указаний на объекты формальных контекстов вида (5.1) отдельных $gth \in Gth$;
- M_7 — множество связей «основа–флексия» для синтаксически зависимого слова;
- M_8 — множество сочетаний основ зависимого и главного слова.

Пусть $Ke = (Ge, Me, Ie)$ есть ФК (5.1) СЯУ S_1 заведомо корректного описания некоторого факта, $Kx = (Gx, Mx, Ix)$ — ФК вида (5.1) для произвольной СЯУ S_2, а $M_U = \left(\bigcup_{i=6}^{8} M_i\right) \cup Me_4 \cup Mx_4 \cup Me_5 \cup Mx_5$.

p_{fl} и p_b есть обозначения для констант «флексия:» и «основа:».

Простейший случай *схожести* S_1 и S_2: для $\forall gx \in Gx\ \exists ge \in Ge\colon\ gx = ge$ и любой признак $me \in Me$ объекта ge относится к gx.

Пример объекта отдельной СЯУ в формальном контексте тезауруса



Случай 2 соответствия объектов

$gx = ge$, условие *простейшего случая* не выполняется, но существует объект $gth \in Gth$, обладающий признаком $mth_1 \in M_6 : mth_1 = p_b \odot ge$ при обязательном выполнении следующих условий:

$$(\exists\, me_{fl} \in Me_5 : me_{fl} = p_{fl} \odot fe) \rightarrow \rightarrow (\exists\, mth_{17} \in M_7 : mth_{17} = ge \odot «:» \odot fe),$$

при этом $\Big(Ie\,(ge, me_{fl}) \wedge Ix\,(ge, me_{fl})\Big) \rightarrow Ith\,(gth, mth_{17})$;

$$(\exists\, me_{bs} \in Me_1 : me_{bs} = p_{bs} \odot be) \rightarrow (\exists\, mth_{18} \in M_8 : mth_{18} = ge \odot «:» \odot be),$$

при этом $Ie\,(ge, me_{bs}) \rightarrow Ith\,(gth, mth_{18})$;

$$(\exists\, mx_{bs} \in Mx_1 : mx_{bs} = p_{bs} \odot bx) \rightarrow (\exists\, mth_{28} \in M_8 : mth_{28} = ge \odot «:» \odot bx),$$

при этом $Ix\,(ge, mx_{bs}) \rightarrow Ith\,(gth, mth_{28})$.

Кроме того, для $\forall\, mth \in (Mth \setminus M_U)$ верно:

$$Ith\,(gth, mth) \rightarrow \Big(Ie\,(ge, mth) \wedge Ix\,(ge, mth)\Big). \qquad (5.3)$$

Случай 3 соответствия объектов

$gx \neq ge$, но существует объект $gth \in Gth$, обладающий признаками

$$mth_1 \in M_6 : mth_1 = p_b \odot ge \text{ и}$$
$$mth_2 \in M_6 : mth_2 = p_b \odot gx,$$

при этом для любого признака $mth \in (Mth \backslash M_U)$ справедливо:

$$Ith\,(gth, mth) \rightarrow \Big(Ie\,(ge, mth) \wedge Ix\,(gx, mth)\Big). \tag{5.4}$$

Замечание

Численная оценка схожести СЯУ включает сравнение последовательностей двух и более соподчинённых слов. Случаи схожести здесь анализируются только для главных слов. Последовательности считаются заменяемыми, если возможно их построение по формальному контексту (5.2) на наборе признаков с префиксом p_{bs} для одной и той же СЯУ.

Случай 4 соответствия объектов

$gx \neq ge$, но существует $gth_1 \in Gth$, обладающий признаком $mth_1 \in M_6$: $mth_1 = p_b \odot ge$, а для $\forall\, me \in (Me_4 \cup Me_5)$ верно то, что

$$\Big(Ith\,(gth_1, mth_1) \wedge Ie\,(ge, me)\Big) \rightarrow Ith\,(gth_1, me)\,.$$

При этом существуют признаки $mth_2 \in M_6$: $mth_2 = p_b \odot gxg$ и $mx \in (Mx_1 \cup Mx_2 \cup Mx_3)$, для которых верно:

$$\Big(Ith\,(gth_1, mth_2) \wedge Ix\,(gx, mx)\Big) \rightarrow Ith\,(gth_1, mx)\,,$$

где $gxg \neq gx$, а пара (gxg, ge) соответствует *Случаю 3* при генерации формального контекста вида (5.1) для gth_1.

В то же время существует объект $gth_2 \in Gth$, относительно которого пара (gx, gxg) также будет соответствовать *Случаю 3*.

Генерируемый при этом формальный контекст вида (5.1) для gth_2 обозначим далее как Kxg, $Kxg = (Gxg, Mxg, Ixg)$.

Оценка схожести ситуаций языкового употребления S_1 и S_2 относительно их формальных контекстов $Ke = (Ge, Me, Ie)$ и $Kx = (Gx, Mx, Ix)$ вычисляется по формуле:

$$spc\,(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{5.5}$$

где $n = |Gx|$,

spc_k есть численное значение схожести объектов в паре (gx_k, ge).

Если (gx_k, ge) не относится ни к одному из четырёх случаев соответствия объектов схожих СЯУ, то $spc\,(S_1, S_2) = 0$.

При взаимно-однозначном соответствии признаков объектов ge и gx значение spc_k равно 1.0.

Если пара (gx_k, ge) отвечает одному из *Случаев 2–3* соответствия объектов, то оценка схожести gx_k и ge вычисляется по формуле:

$$-\log_2\left(1-\frac{D_c}{path_C}\right)\times\frac{\left|BLCS\right|}{\left|B_1\backslash BLCS\right|+\left|B_2\backslash BLCS\right|+\left|BLCS\right|}, \qquad (5.6)$$

где $D_c = 2$, число $path_C = 4$.

В множество $BLCS$ войдут признаки $mth \in (Mth\backslash M_U)$, для которых справедливо либо соотношение (5.3) для *Случая 2*,
либо соотношение (5.4) для *Случая 3*.

При этом

$$B_1 = \left\{ me : me \in \left(Me_1 \cup Me_2 \cup Me_3 \right), Ie(ge, me) = \text{true} \right\},$$
$$B_2 = \left\{ mx : mx \in \left(Mx_1 \cup Mx_2 \cup Mx_3 \right), Ix(gx_k, mx) = \text{true} \right\}.$$

Если пара (gx_k, ge) отвечает *Случаю 4* соответствия объектов, то оценка схожести gx_k и ge вычисляется по формуле:

$$-\log_2\left(1-\frac{D_c}{path_C}\right)\times\frac{\left|BLCS\right|}{\left|B_1\backslash BLCS\right|+\left|B_2\backslash BLCS\right|+\left|BLCS\right|}. \qquad (5.6)$$

Для рассматриваемого случая имеем:

$$B_1=\left\{\, mxg\colon mxg\in\left(\, Mxg_1\cup Mxg_2\cup Mxg_3\,\right), Ixg\left(gxg, mxg\right)=\text{true}\,\right\},$$

$$B_2=\left\{\, mx\colon mx\in\left(\, Mxg_1\cup Mxg_2\cup Mxg_3\,\right), Ixg\left(gx_k, mx\right)=\text{true}\,\right\},$$

где $D_c=2$, $\left(Mxg_1\cup Mxg_2\cup Mxg_3\right)\subset Mxg$, $BLCS=B_1\cap B_2$.

Соответствие *Случаю 4* обычно проверяется в несколько итераций.

В ходе каждой последующей итерации число признаков, не являющихся общими для gx_k и gxg, всегда меньше, чем в предыдущей.

Начальное значение $path_C=4$ и с каждым шагом возрастает на 1.

Исходные данные для построения фрагмента тезауруса

№п/п	1				2	3		4	
основа	флективная часть + предлог								
заниженн	ость	ость	ости	ости	—	ость	ости	ость	ость
оценк	—	—	—	—	—	и	и	и	и
эмпирическ	ого	—	ого	—	—	—	—	—	—
риск	а	—	а	—	—	—	—	—	—
средн	—	ей	—	ей	—	—	—	—	—
ошибк	—	и:на	—	и:на	—	—	—	и	и
распознавани	—	—	—	—	—	—	—	я	я
обучающ	—	ей	—	ей	—	—	—	—	—
выборк	—	е	—	е	—	—	—	—	—
переусложнени	ем	ем	е	е	—	—	—	—	—
модел	и	и	и	и	—	—	—	—	—
уменьшени	—	—	—	—	е	—	—	—	—
обобщающ	—	—	—	—	ей	ей	ей	—	—
способност	—	—	—	—	и	и	и	—	—
выбор	—	—	—	—	—	—	—	ом	а
решающ	—	—	—	—	его	—	—	его	его
дерев	—	—	—	—	а	—	—	—	—
правил	—	—	—	—	—	—	—	а	а
алгоритм	—	—	—	—	—	а	а	—	—
переподгонк	—	—	—	—	ой	ой	а	—	—
переобучени	—	—	—	—	—	ем	е	—	—
связан	а:с	а:с	—	—	о:с	а:с	—	а:с	—
вызван	а	а	—	—	—	а	—	—	—
обусловлен	а	а	—	—	о	—	—	—	—
привод	—	—	ит:к	ит:к	—	—	ит:к	—	—
завис	—	—	—	—	—	—	—	—	ит:от

Численная оценка схожести ответа с эталоном

ответы	эталон				анализируемый		
вариант	1	2	3	4	1	2	3
основа	флективная часть + предлог						
заниженн	ости	ости	ость	ость	ость	ость	ости
эмпирическ	ого	ого	ого	ого	—	—	—
риск	а	а	а	а	—	—	—
средн	—	—	—	—	ей	ей	ей
ошибк	—	—	—	—	и:на	и:на	и:на
обучающ	—	—	—	—	ей	ей	ей
выборк	—	—	—	—	е	е	е
переобучени	е	—	—	ем	ем	—	е
переподгонк	—	а	ой	—	—	ой	—
связан	—	—	а:с	а:с	а:с	а:с	—
привод	ит:к	ит:к	—	—	—	—	ит:к

Вариант	$spc(S_1, S_2)$	$\|BLCS\|$	$\|B_1 \backslash BLCS\|$	$\|B_2 \backslash BLCS\|$
1	0.9167	7.7500	0.7500	0.0000
2	0.7917	7.0000	2.0000	0.5000
3	0.8750	7.7500	0.7500	0.7500

Основные задачи и результаты шестой главы

Основные задачи:

1. Разделение и сжатие баз предметных и языковых знаний с применением комплексной методики формирования и кластеризации семантических отношений, последовательно развиваемой в третьей, четвёртой и пятой главах.
2. Разработка архитектуры программной системы контроля знаний, реализующей предложенные в работе принципы, методы и алгоритмы.

Важнейший результаты (п. 6 паспорта специльности 05. 13. 17):

1. Метод и алгоритмы автоматизированного формирования смыслового эталона, а также метод компрессии текстовой базы знаний на основе выделенных эталонов.
2. Демо-версия программной системы контроля знаний на основе тестовых заданий открытой формы (имеется акт об опытной эксплуатации системы в НИИПТ «Растр», г. Великий Новгород, а также акт об апробации результатов НИР в учебном процессе Новгородского государственного университета).

Пусть согласно определению 3.5 (V_{JT}, I_{JT}) есть расширенное дерево вида (3.11) для множества СЭ-фраз, $I_{JT} = \{(j,k) \colon \exists E \in V_J, (j,k) \in E\}$,

V_J — множество вершин графа синтагм,

$Ke = (Ge, Me, Ie)$ есть искомый формальный контекст эталона.

Если $\exists\ E \in V_J \colon\ (j,k) \in E$ в дереве (V_{JT}, I_{JT}), то для основ b_j и b_k и флексий f_j и f_k элементы множеств Ge, Me и отношения Ie формируются следующим образом.

Случай 1

Индекс k соответствует родительскому узлу, j — дочернему, линейная структура фразы не содержит предлог между словами с индексами j и k.

При этом в Me включаются признаки $m_1 = p_{bs} \odot b_k$, $m_2 = p_{bf} \odot f_k$, $m_3 = p_{fl} \odot f_j$ и $m_4 = f_j \odot \text{«:»} \odot f_k$, основа b_j включается в множество Ge, $Ie = Ie \cup \{(b_j, m_1),\ (b_j, m_2),\ (b_j, m_3),\ (b_j, m_4)\}$.

Случай 2

Между словами с индексами j и k стоит предлог p_y.

Ie, m_1 и m_3 — аналогично *Случаю 1*, $m_2 = p_{bf} \odot f_k \odot \text{«:»} \odot p_y$, $m_4 = f_j \odot \text{«:»} \odot f_k \odot \text{«:»} \odot p_y$.

Коэффициент сжатия по основам при представлении СЯУ формальным контекстом вида (5.1) определяется как

$$ks = \frac{\sum_{i=1}^{nbs} ks_i}{nbs}, \tag{6.4}$$

где в соответствии с принятым разбиением множества Ms

$$ks_i = \frac{\sum_{j=1}^{nbs_i} \sum_{k=1}^{nmf} nas_{ijk}}{nbs_i};$$

$$nbs = \left|M_1\right|;$$

$$nmf = \left|M_2\right|;$$

$$nbs_i = \left|\left\{ g \in Gs \colon Is\,(g, m) = \text{true},\ m \in M_1,\ m = p_{bs} \odot b_i \right\}\right|;$$

$$nas_{ijk} = \left|\left\{ m_k \in M_3 \colon Is\,(g_j, m_k) = \text{true},\ \exists\, m_{bf} \in M_2,\ m_{bf} = p_{bf} \odot f_k,\ m_k = b_i \odot «:» \odot f_k \right\}\right|;$$

p_{bf} соответствует символьной константе «главное-флексия:».

Коэффициент сжатия по флексиям аналогичен коэффициенту по основам:

$$kf = \frac{\sum_{i=1}^{nfs} kf_i}{nfs}, \tag{6.5}$$

где

$$kf_i = \frac{\sum_{j=1}^{nfs_i} \sum_{k=1}^{nmf} naf_{ijk}}{nfs_i};$$

$$nfs = \left|M_5\right|;$$

$$nmf = \left|M_2\right|;$$

$$nfs_i = \left|\left\{ g \in Gs \colon Is\,(g, m) = \text{true},\ m \in M_5,\ m = p_{fl} \odot f_i \right\}\right|;$$

$$naf_{ijk} = \left|\left\{ m \in M_4 \colon Is\,(g_j, m) = \text{true},\ \exists\ m_{bf} \in M_2,\ m_{bf} = p_{bf} \odot f_k,\ m = f_i \odot «:» \odot f_k \right\}\right|;$$

p_{bf} соответствует символьной константе «главное-флексия:».

Производится по максимуму сжатия информации по основам и флексиям в результирующем формальном контексте.

Утверждение

Признак из состава множества признаков формального контекста фразы может быть включён в множество признаков формального контекста эталона, если он входит в признаковую пятёрку $\{m_1, m_2, m_3, m_4, m_5\}$, в которой $m_1 = p_{bs} \odot b$, $m_2 = p_{bf} \odot f_1$, $m_3 = b \odot «:» \odot f_1$, $m_4 = p_{fl} \odot f_2$, $m_5 = f_2 \odot «:» \odot f_1$, а b есть основа некоторого слова. При этом основе b не должен соответствовать объект формального контекста, если есть другой объект этого же контекста, который обладает одновременно признаком m_1 и некоторым другим признаком $m = p_{bs} \odot b_1$, где $b_1 \neq b$, а основе b_1 не соответствует ни одного объекта этого формального контекста при том, что признак m относится более чем к одному объекту.

Замечание

Последовательности из трех и более соподчиненных слов, встречающиеся более чем в 49% исходных СЭ-фраз, выделяются на этапе синтаксического разбора. Для каждой из них строится отдельный формальный контекст, идентичный по структуре формальному контексту эталона.

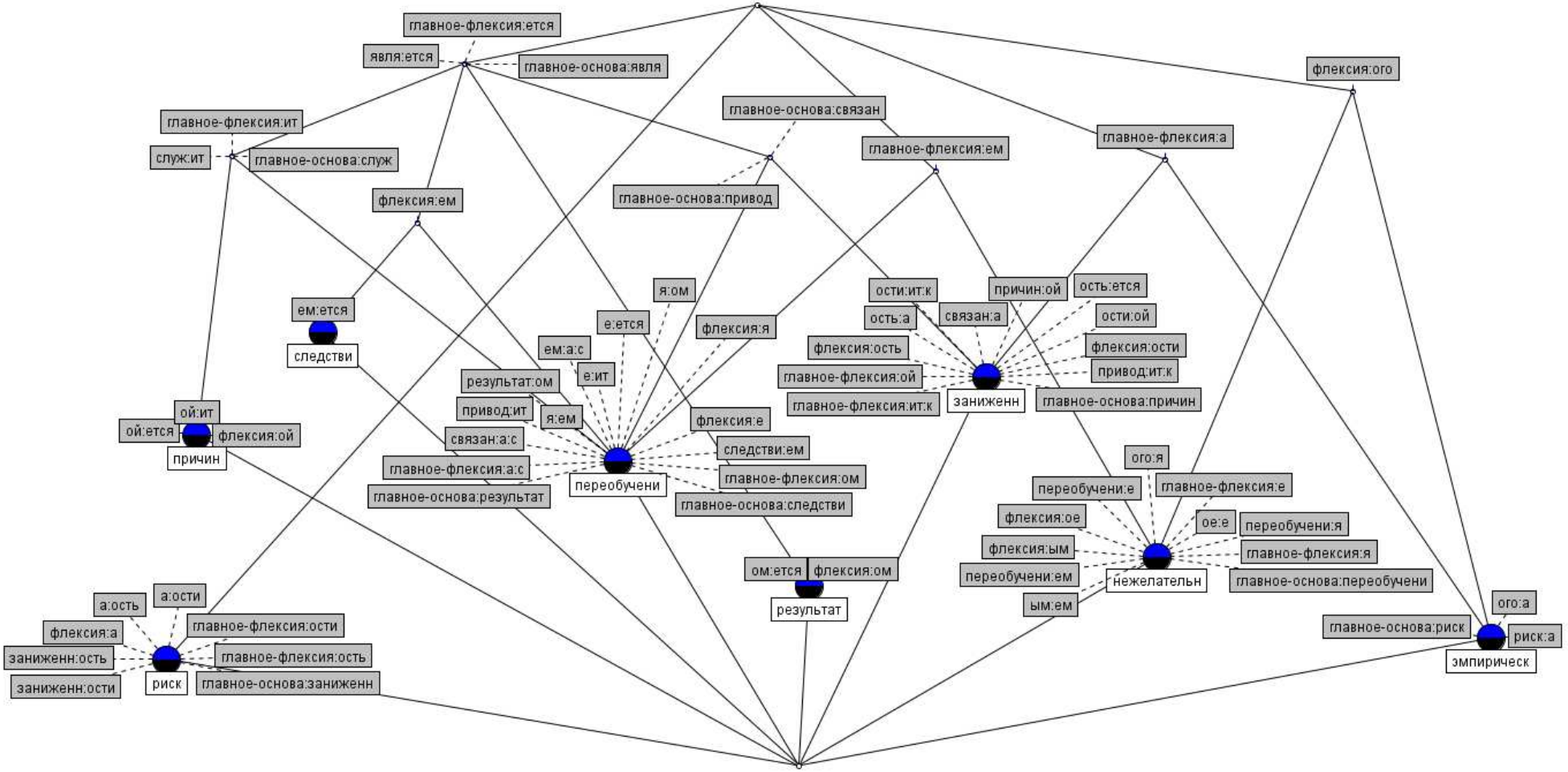
Пример: исходное множество семантически эквивалентных фраз

Синонимичные перифразы

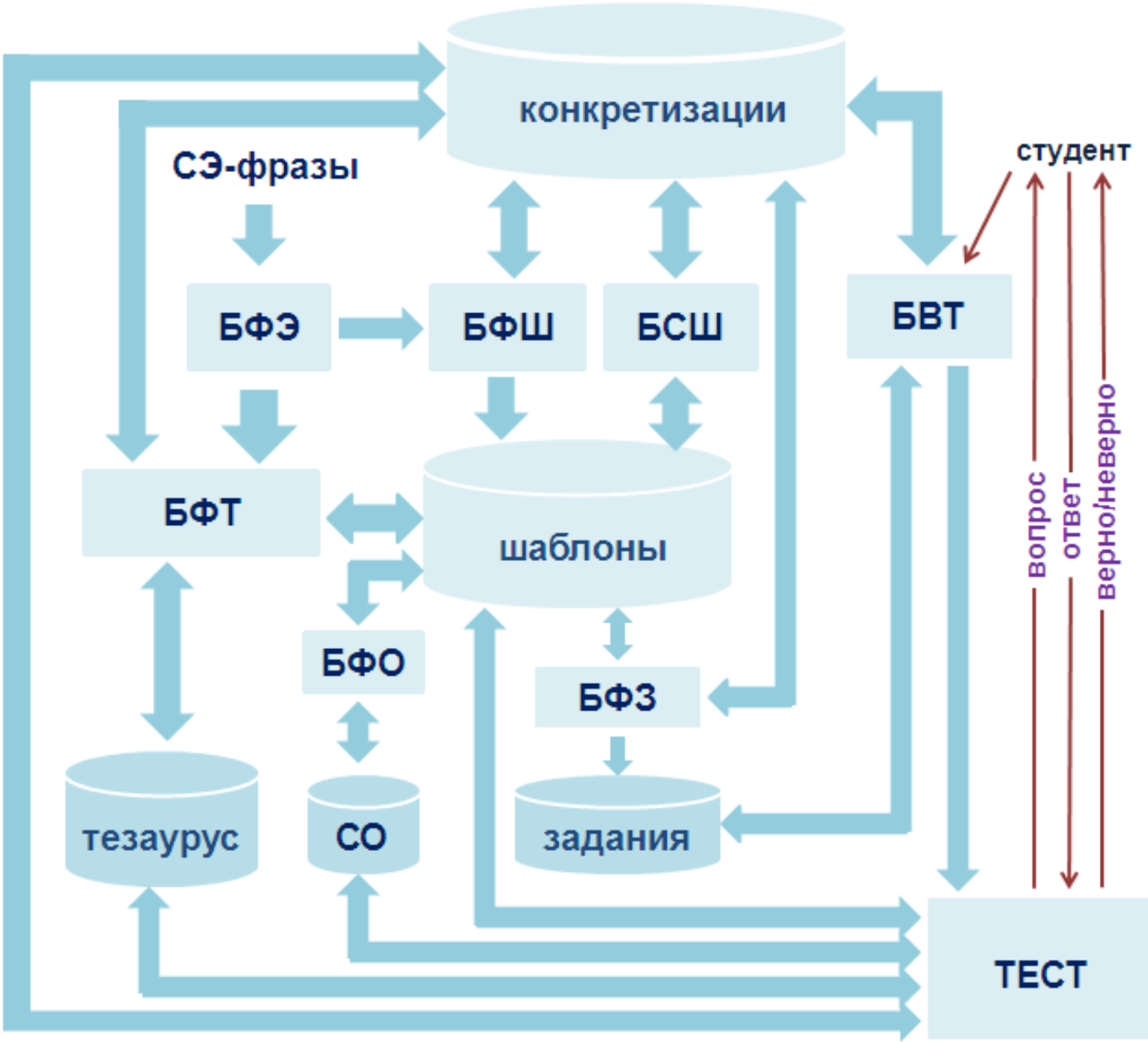
27:89 Insert Indent Modified

"Нежелательное переобучение приводит к заниженности эмпирического риска.",
"Нежелательное переобучение, следствием которого является заниженность эмпирического риска.",
"Заниженность эмпирического риска является следствием нежелательного переобучения.",
"Заниженность эмпирического риска, являющаяся следствием нежелательного переобучения.",
"Эмпирический риск, заниженность которого является следствием нежелательного переобучения.",
"Эмпирический риск, заниженный вследствие нежелательного переобучения.",
"Эмпирический риск, к заниженности которого ведет нежелательное переобучение.",
"Риск, заниженный как следствие переобучения.",
"Эмпирический риск по причине, обусловленной нежелательным переобучением, может оказаться заниженным. ",
"Эмпирический риск в силу обстоятельств, связанных с нежелательным переобучением, может оказаться заниженным.
"Эмпирический риск по причине, вызванной нежелательным переобучением, может быть заниженным. ",
"Эмпирический риск, к заниженности которого приводит нежелательное переобучение.",
"Нежелательное переобучение служит причиной заниженности эмпирического риска.",
"Заниженность эмпирического риска, причиной которой является нежелательное переобучение.",
"Заниженность эмпирического риска является результатом нежелательного переобучения.",
"Нежелательное переобучение, с которым связана заниженность эмпирического риска.",
"Эмпирический риск, с переобучением связана его заниженность. ",
"Заниженность эмпирического риска связана с переобучением.",
"Заниженность эмпирического риска, являющаяся результатом нежелательного переобучения.",
"Нежелательное переобучение, результатом которого является заниженность эмпирического риска.",
"Нежелательное переобучение, результат которого есть заниженность эмпирического риска.",
"Нежелательное переобучение, приводящее к заниженности эмпирического риска.",
"Нежелательное переобучение, служащее причиной заниженности эмпирического риска.",
"Заниженность эмпирического риска относится к следствию нежелательного переобучения.",
"Заниженность эмпирического риска связана с нежелательным переобучением.",
"Нежелательное переобучение является причиной заниженности эмпирического риска.",
"Заниженность эмпирического риска, причиной которой служит нежелательное переобучение."

Результат: формальный контекст смыслового эталона



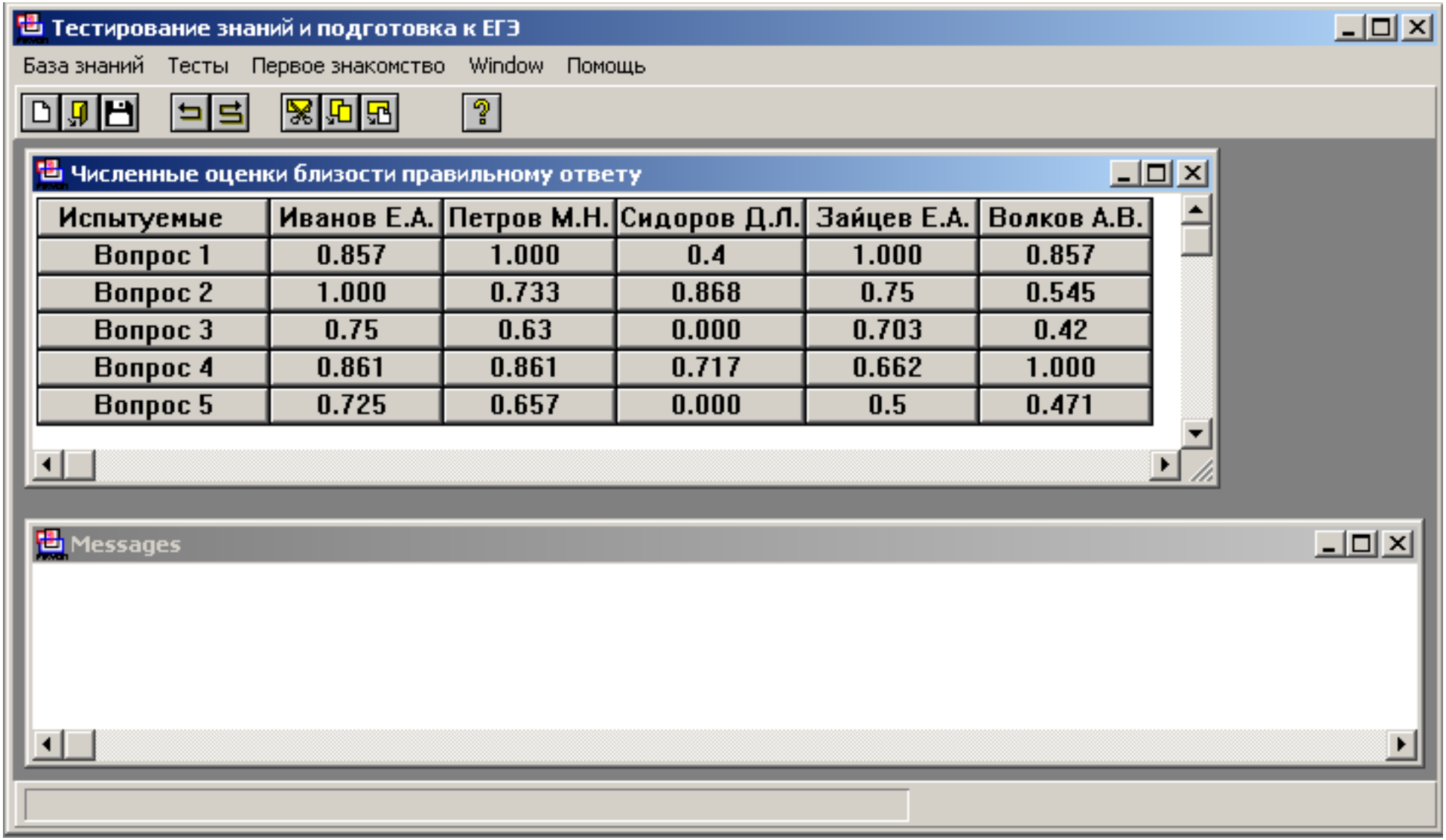
Архитектура программной системы тестирования знаний



Обозначения:

БФЭ — блок формирования эталонов;
БФШ — блок формирования шаблонов;
БСШ — блок слияния шаблонов;
БФТ — блок формирования тезауруса;
БВТ — блок выбора теста;
ТЕСТ — блок выполнения теста;
БФЗ — блок формирования заданий;
СО — база синтаксических отношений;
БФО — блок формирования отношений.

Интерфейс системы



Испытуемые	Иванов Е.А.	Петров М.Н.	Сидоров Д.Л.	Зайцев Е.А.	Волков А.В.
Вопрос 1	0.857	1.000	0.4	1.000	0.857
Вопрос 2	1.000	0.733	0.868	0.75	0.545
Вопрос 3	0.75	0.63	0.000	0.703	0.42
Вопрос 4	0.861	0.861	0.717	0.662	1.000
Вопрос 5	0.725	0.657	0.000	0.5	0.471

Демо-версия системы представлена на www.machinelearning.ru, на персональной странице автора.

Возможные случаи ответа и их интерпретация

Случай 1. Неполный ответ — все слова и словосочетания из ответа испытуемого нашли прообразы в наиболее близком варианте правильного ответа, но часть слов правильного ответа не нашла прообразов в ответе испытуемого.

Ненулевое значение оценки (5.6) будет для упущенного слова, которое в правильном ответе синтаксически зависимо относительно некоторого другого слова, присутствующего в анализируемом ответе.

Значение оценки (5.6) для упущенного слова здесь равно

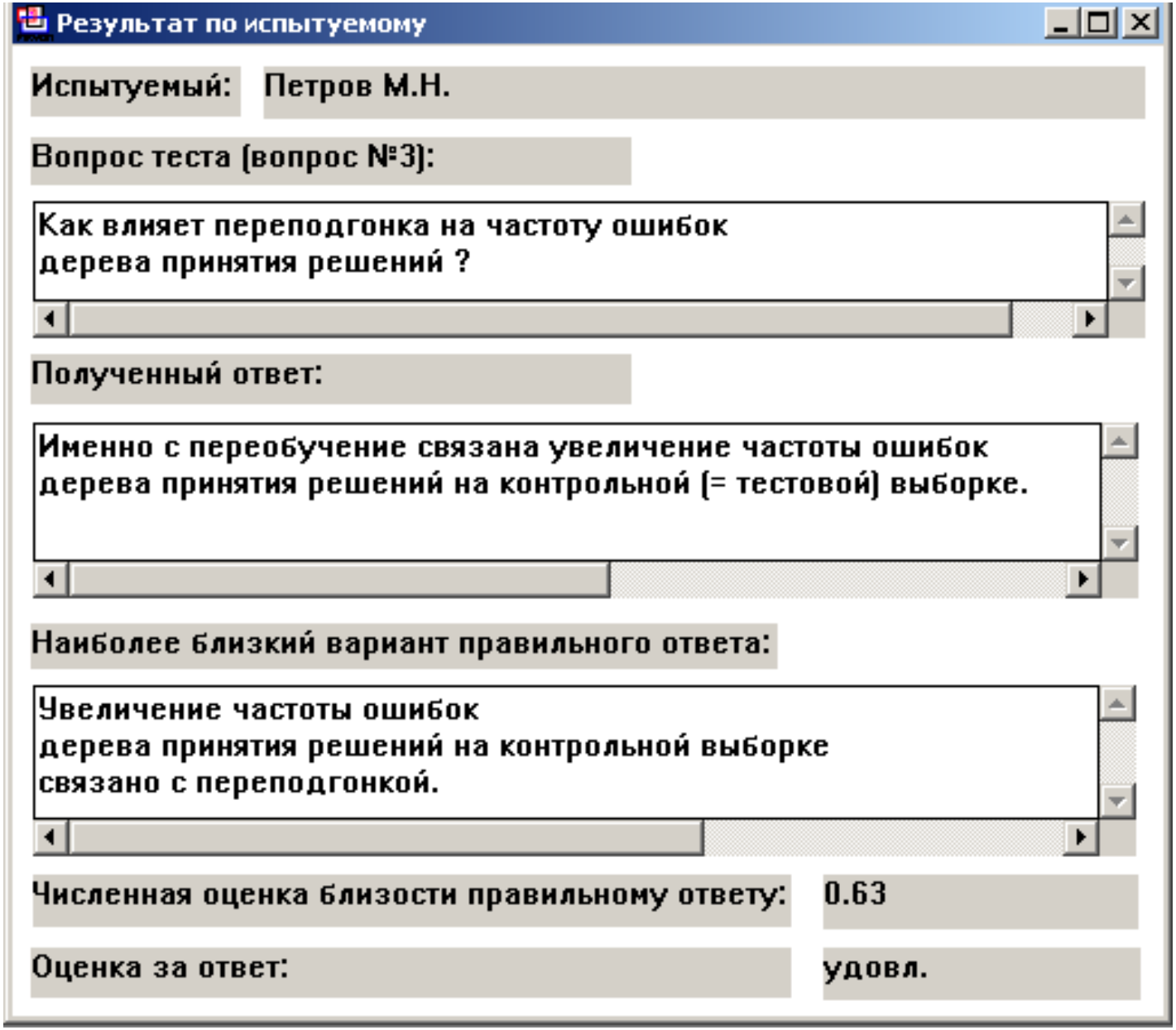
$$-\log_2\left(1-\frac{2}{4}\right)\times\frac{3}{(8-3)+(8-3)+3}\approx 0.23.$$

Случай 2. Орфографические ошибки (из допустимых) — слово из ответа испытуемого и слово правильного ответа есть формы одного и того же слова в рамках некоторой лексико-синтаксической связи из известных системе.

Случай 3. «Лишние» слова — в анализируемом ответе имеются слова, не нашедшие прообразов в правильном «варианте».

Ответ засчитывается как неверный, если слова фигурируют в лексико-синтаксических связях из известных системе.

Пример выдачи результатов по отдельному испытуемому

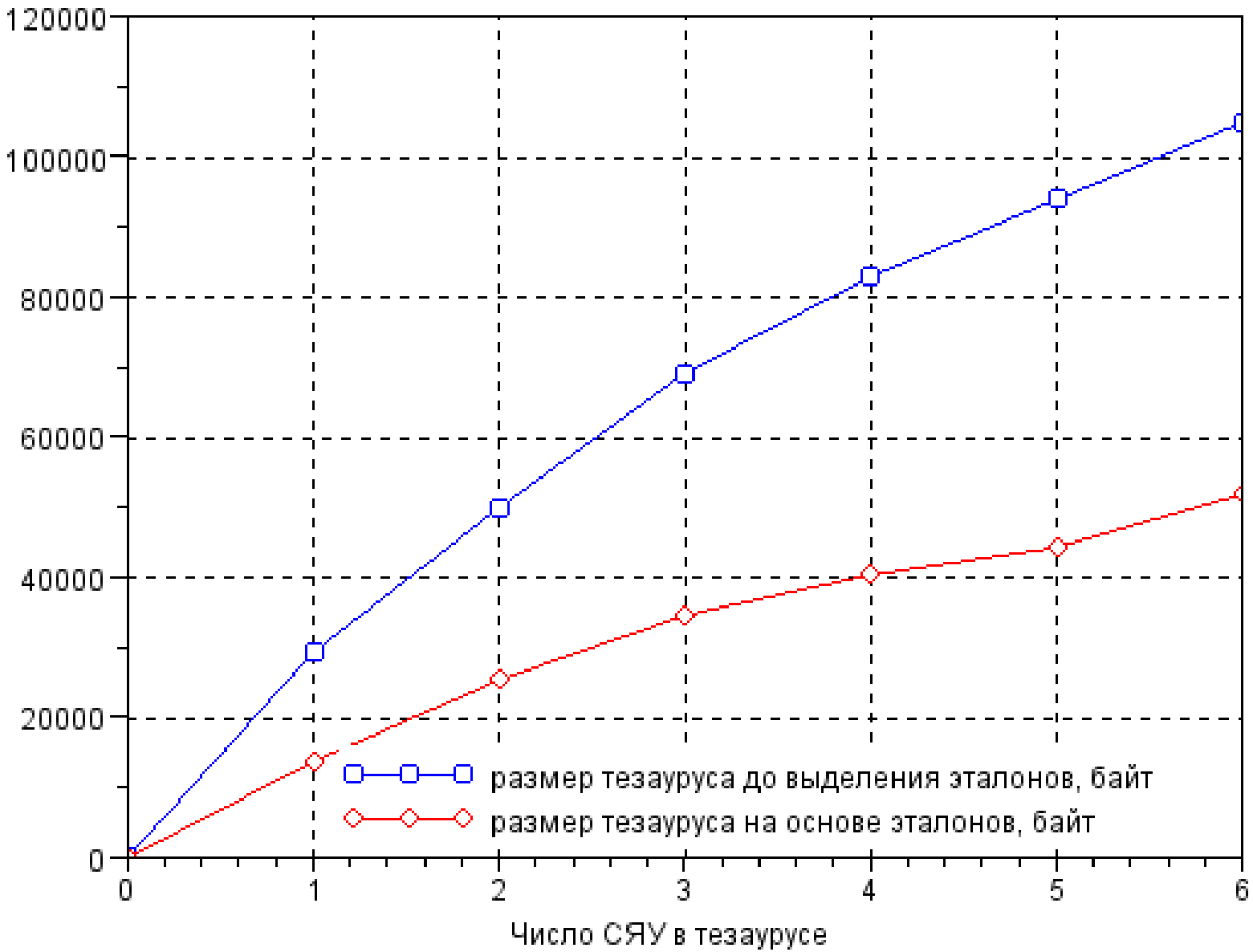


Ситуации языкового употребления и смысловые эталоны

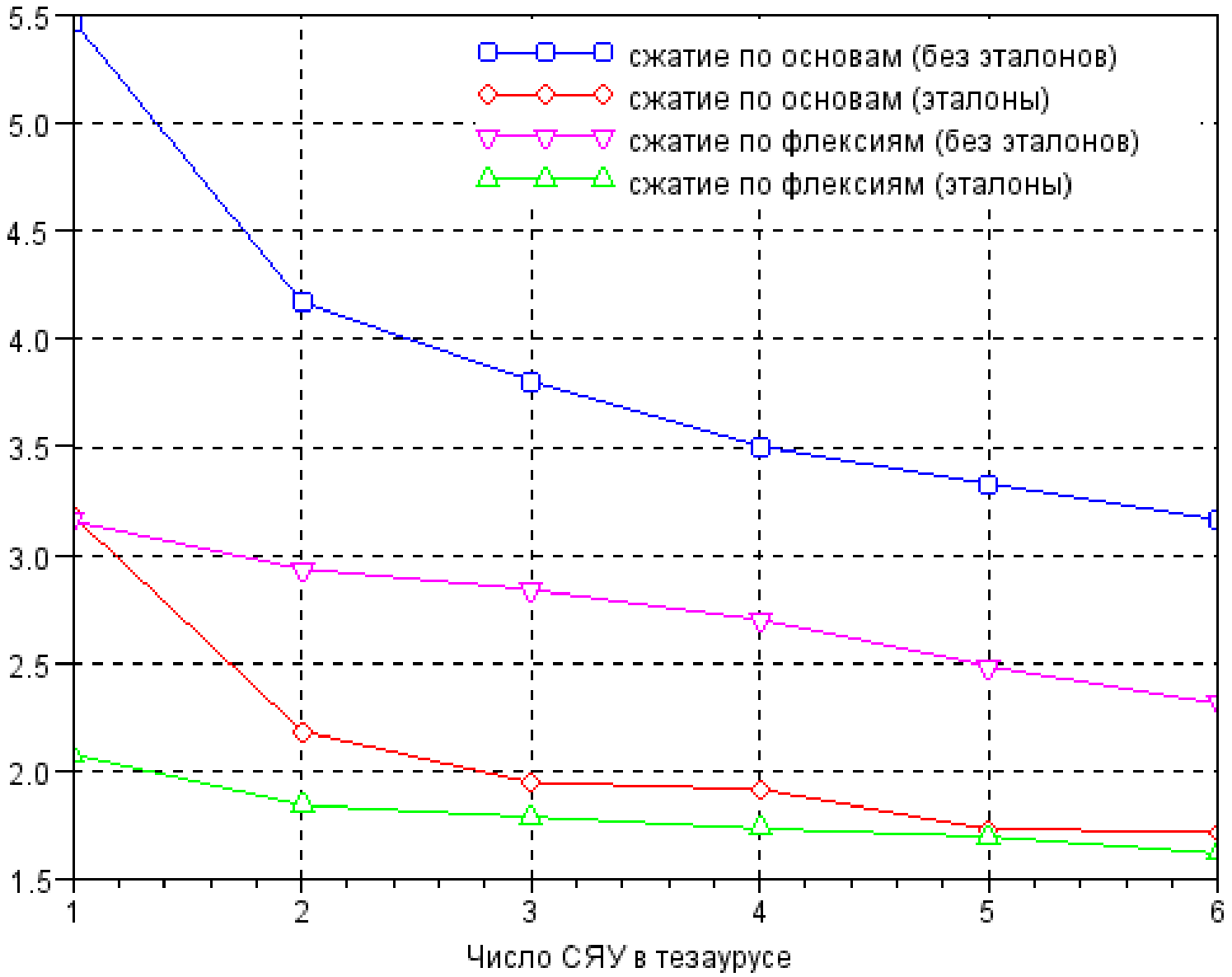
Порядковый номер СЯУ, i	1	2	3	4	5	6
Число фраз, задающих СЯУ	56	28	29	30	6	10
из них представляют эталон	8	9	7	9	1	2
Исходное число объектов СЯУ	18	17	15	13	12	14
Исходное число признаков СЯУ	177	186	173	162	94	81
Число объектов эталона	9	12	12	11	8	12
Число признаков эталона	82	90	80	69	35	53

i	Ситуация языкового употребления
1	Связь переобучения с эмпирическим риском
2	Связь переусложнения модели с заниженностью средней ошибки на тренировочной выборке
3	Влияние переподгонки на частоту ошибок дерева принятия решений
4	Причина заниженности оценки обобщающей способности алгоритма
5	Зависимость оценки ошибки распознавания от выбора решающего правила
6	Зависимость обобщающей способности логического алгоритма классификации от числа закономерностей алгоритмической композиции

Соотношение размеров тезауруса для разного числа СЯУ



Сжатие информации относительно формального контекста тезауруса



Оценка требуемого объёма памяти для хранения ЕЯ-фразы

i	1	2	3	4	5	6
n	12	15	16	17	10	14
$vol\,(n)$	$4.790 \cdot 10^8$	$1.308 \cdot 10^{12}$	$2.092 \cdot 10^{13}$	$3.557 \cdot 10^{14}$	$3.629 \cdot 10^6$	$8.718 \cdot 10^{10}$
$vol_1\,(n)$	648	795	416	442	20	42
$vol_2\,(n)$	168	225	80	187	20	42

Здесь:

i — порядковый номер СЯУ;
n — максимальное число слов во фразе;
$vol\,(n) = n!$ есть традиционно используемая оценка;
vol_1 и vol_2 — оценки с применением метода и алгоритмов выделения эталона СЯУ, предложенных в диссертации.

При этом:

$vol_1\,(n) = l_1 \cdot n$ есть оценка сверху, l_1 — число фраз, определяющих СЯУ;
$vol_2\,(n) = l_2 \cdot n$ есть оценка снизу, l_2 — число фраз, определяющих эталон СЯУ.

Согласование знаний о синонимии относительно разных ситуаций языкового употребления

Пусть

St — основа слова (его неизменная часть);
Fl — его флексия;
S_1 и S_2 — некоторые СЯУ.

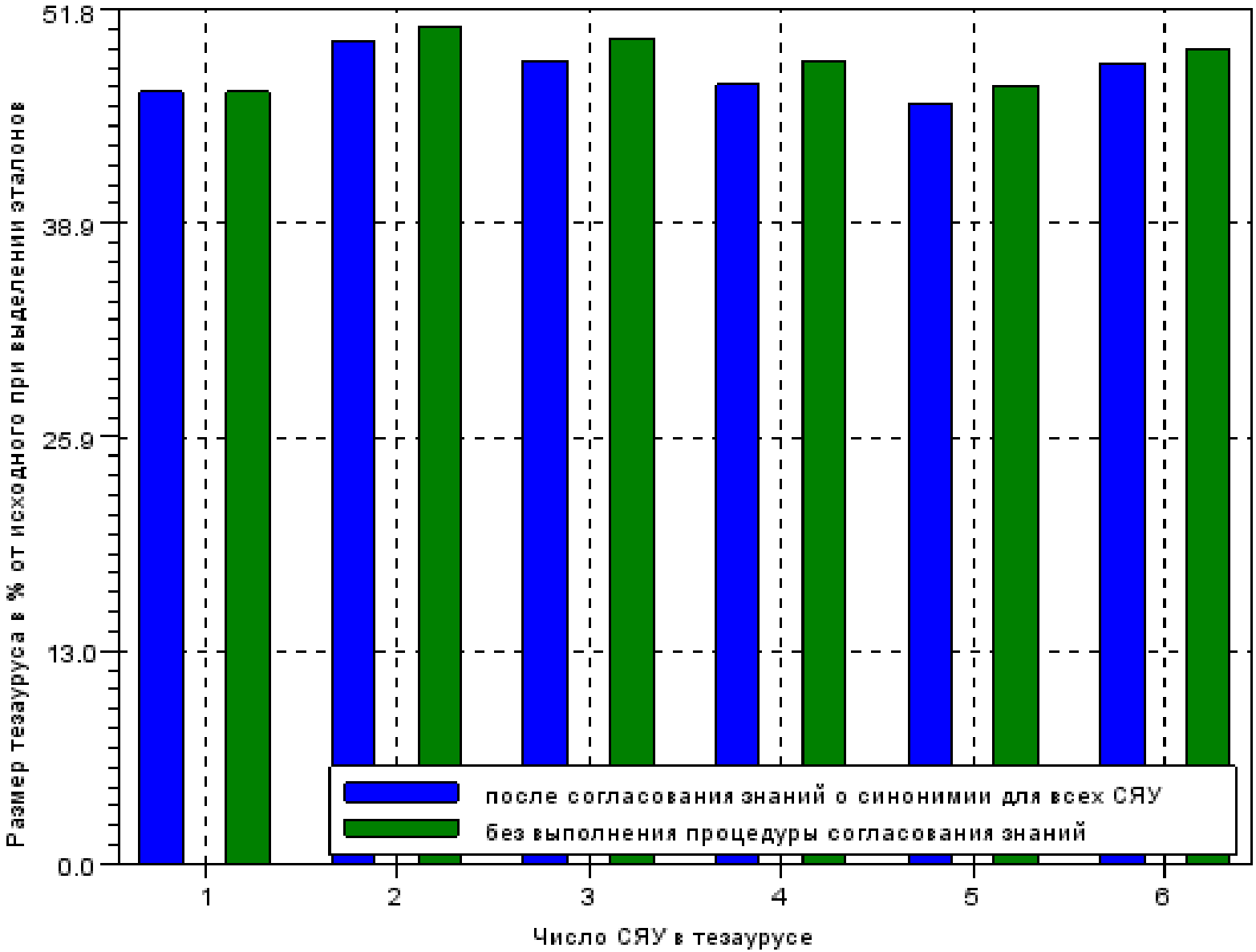
Предположим, что некоторое слово Wrd относительно S_1 представляется как $St_1 \odot Fl_1$, а относительно S_2 — как $St_2 \odot Fl_2$, причём $St_1 = St_2 \odot Sf$, где Sf содержит минимум один символ, а $\odot$ есть операция конкатенации символьных строк.

Тогда относительно S_1 основа St_1 будет заменена на St_2, флексия Fl_1 заменяется на $Fl_3 = Sf \odot Fl_2$, но только в том случае, если частоты встречаемости флексий Fl_3 и Fl_2 во всех лексико-синтаксических связях, представляемых формальным контекстом вида (5.2) для заданной предметной области, не уменьшаются при выполнении указанных замен.

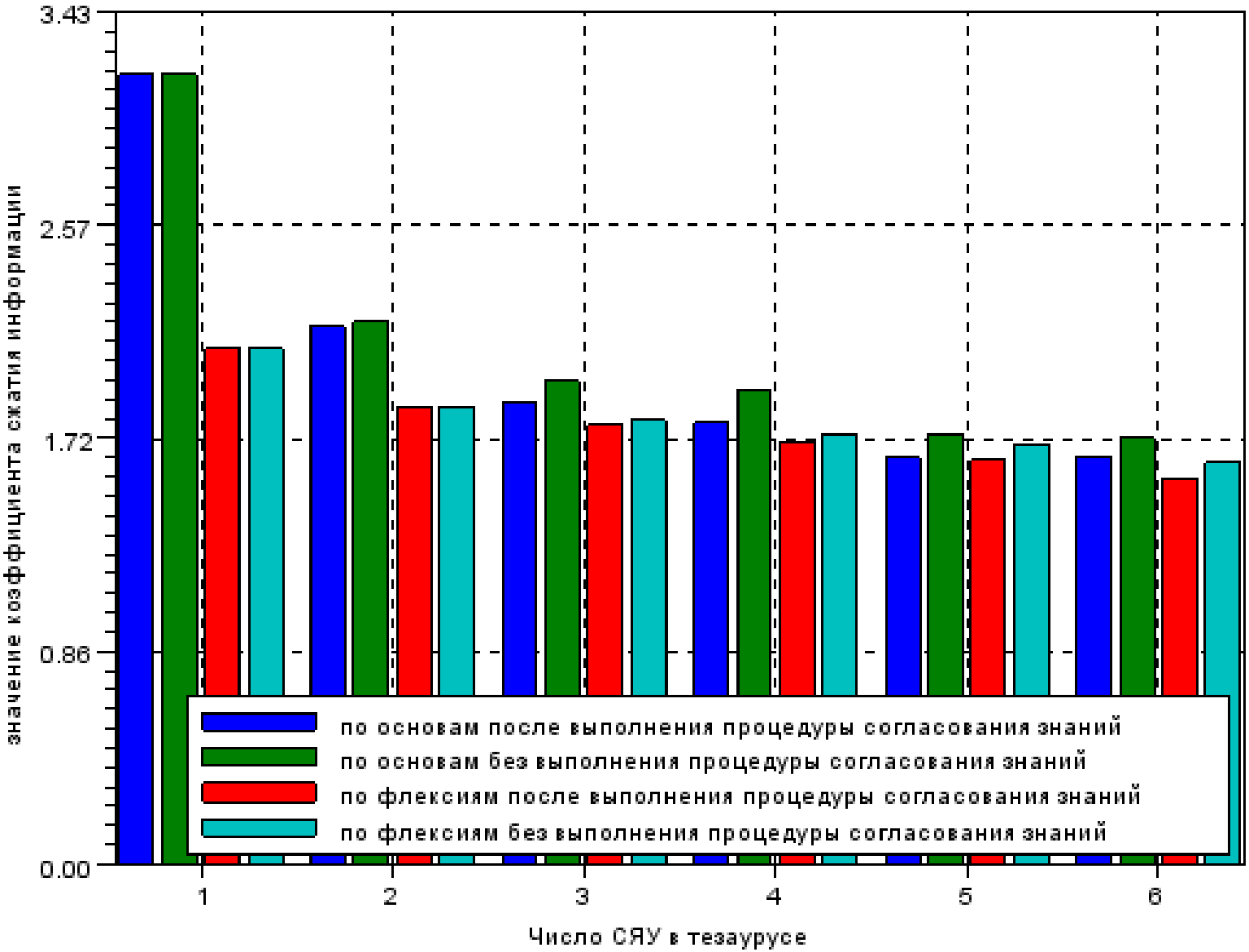
Пример.

СЯУ №3, $St_1 =$ «является», $Fl_1 =$ «»,
СЯУ №1, $St_2 =$ «явля», $Fl_2 =$ «ется», $Sf =$ «ется»,
а относительно СЯУ №3 производится замена: Fl_1 — на $Fl_3 =$ «ется».

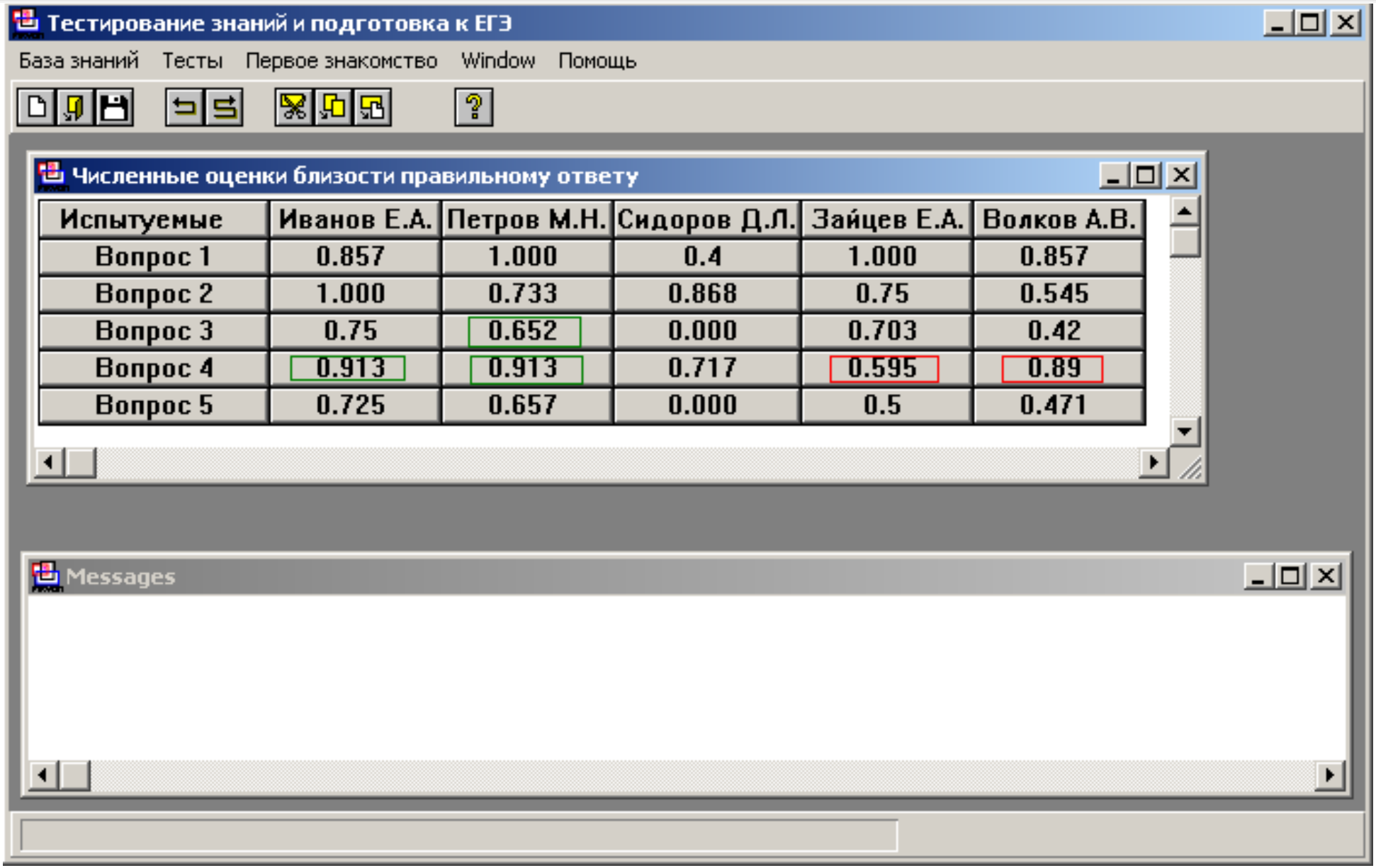
Сокращение размеров тезауруса согласованием знаний о синонимии



Относительность знаний о синонимии и сжатие по основам/флексиям



Результаты группового тестирования после автоматического согласования знаний о синонимии относительно различных СЯУ



Испытуемые	Иванов Е.А.	Петров М.Н.	Сидоров Д.Л.	Зайцев Е.А.	Волков А.В.
Вопрос 1	0.857	1.000	0.4	1.000	0.857
Вопрос 2	1.000	0.733	0.868	0.75	0.545
Вопрос 3	0.75	0.652	0.000	0.703	0.42
Вопрос 4	0.913	0.913	0.717	0.595	0.89
Вопрос 5	0.725	0.657	0.000	0.5	0.471

Основные результаты, выносимые на защиту

1. Комплексная методика автоматизированного формирования и экспериментальной оценки знаний в виде классов семантической эквивалентности текстов на основе ситуаций языкового употребления.
2. Информационная модель совокупности правил грамматики деревьев для выделения сверхфразовых единств и классов семантической эквивалентности на уровне глубинного синтаксиса.
3. Принцип формирования и кластеризации семантических отношений на основе классов семантической эквивалентности.
4. Метод и алгоритмы выделения смыслового эталона на множестве эквивалентных по смыслу фраз предметно-ограниченного естественного языка.
5. Метод компрессии текстовой базы знаний с применением смысловых эталонов.
6. Метод численной оценки семантической схожести текстов предметно-ограниченного естественного языка относительно ситуаций его употребления.

Основные публикации (всего — 75 работ, 18 — в изданиях ВАК)

1. *Михайлов Д. В.* Теоретические основы построения открытых вопросно-ответных систем. Семантическая эквивалентность текстов и модели их распознавания: монография / Д. В. Михайлов, Г. М. Емельянов; НовГУ им. Ярослава Мудрого. Великий Новгород, 2010. 286 с.
2. *Mikhailov D. V.* Recognition of Superphrase Unities in Texts while Establishing Their Semantic Equivalence / G. M. Emelyanov, D. V. Mikhailov, E. I. Zaitseva // Pattern Recognition and Image Analysis. 2003. Vol. 13, No. 3. P. 447–451.
3. *Михайлов Д. В.* Распознавание сверхфразовых единств при установлении эквивалентности смысловых образов высказываний в общей задаче моделирования языковой деятельности / Г. М. Емельянов, Д. В. Михайлов // Известия СПбГЭТУ «ЛЭТИ», сер. «Информатика, управление и компьютерные технологии». СПб., 2003. Вып. 1. С. 65–73.
4. *Михайлов Д. В.* Семантическая кластеризация текстов предметных языков (морфология и синтаксис) / Д. В. Михайлов, Г. М. Емельянов // Компьютерная оптика. 2009. Т. 33, № 4. С. 473–480.
5. *Mikhailov D. V.* Semantic Clustering and Affinity Measure of Subject-Oriented Language Texts / D. V. Mikhailov, G. M. Emelyanov // Pattern Recognition and Image Analysis. 2010. Vol. 20, No. 3. P. 376–385.
6. *Mikhailov D. V.* Sense's Standards and Machine Understanding of Texts in the System for Computer-Aided Testing of Knowledge / G. M. Emelyanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2011. Vol. 21, No. 4. P. 705–719.
7. *Михайлов Д. В.* Формирование смысловых эталонов и интерпретация результатов открытых тестов в системах контроля знаний / Д. В. Михайлов // Вестник Новгородского государственного университета имени Ярослава Мудрого, сер. «Технические науки». 2011. Т. 33, № 65. С. 83–87.

Основные публикации (материалы конференций РАН)

8. *Михайлов Д. В.* Применение аппарата ограниченных сетей Петри для построения динамической модели естественного языка / Г. М. Емельянов, Е. И. Зайцева, Д. В. Михайлов // Междунар. конф. ИОИ-2002: тез. докл. Симферополь, 2002. С. 121–122.
9. *Mikhailov D. V.* Updating of the language knowledge base in the problem of statement's semantic images's equivalence's analysis / G. M. Emelyanov, D. V. Mikhailov // 7th Int. Conf. PRIA-7-2004. Conf. Proc. / SPbETU. St. Petersburg, 2004. Vol. II, P. 462–465.
10. *Михайлов Д. В.* Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности / Д. В. Михайлов, Г. М. Емельянов // 13-я Всерос. конф. ММРО-13. М., 2007. С. 500–503.
11. *Mikhailov D. V.* Formation and clustering of Russian's nouns's contexts within the frameworks of splintered values / D. V. Mikhailov, G. M. Emelyanov // 9th Int. Conf. PRIA-9-2008. Conf. Proc. / NNSU. Nizhni Novgorod, 2008. Vol. 2. P. 39–42.
12. *Михайлов Д. В.* Морфология и синтаксис в задаче семантической кластеризации / Д. В. Михайлов, Г. М. Емельянов // 14-я Всерос. конф. ММРО-14: сб. докл. М., 2009. С. 563–566.
13. *Михайлов Д. В.* Семантическая схожесть текстов в задаче автоматизированного контроля знаний / Д. В. Михайлов, Г. М. Емельянов // 8-я Междунар. конф. ИОИ-2010: сб. докл. М., 2010. С. 516–519.
14. *Михайлов Д. В.* Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний / Г. М. Емельянов, Д. В. Михайлов // 15-я Всерос. конф. ММРО-15: сб. докл. М., 2011. С. 581–584.