Теоретические основы оценки семантической эквивалентности, модели распознавания и компрессии текстов в открытых системах контроля знаний

Михайлов Дмитрий Владимирович

Диссертация на соискание учёной степени
доктора физико-математических наук

05. 13. 18 – математическое моделирование,
численные методы и комплексы программ

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

Великий Новгород, 2011 г.

Цель и задачи исследования

Основная цель диссертационной работы

Разработка теоретико-методологических основ организации обработки Естественного Языка (ЕЯ) для задачи автоматизированного обучения и контроля знаний на основе Тестовых Заданий Открытой Формы (ТЗОФ).

Основные задачи исследования

- Анализ методов моделирования языковых конструкций и определение функциональных требований к механизму сравнения смыслов.
- Разработка и исследование методов моделирования Семантической Эквивалентности (СЭ) на уровне варьирования абстрактной лексикой.
- Разработка методов накопления и систематизации знаний о морфологии и синтаксисе предметно-ограниченного ЕЯ.
- Алгоритмизация механизма оперирования указанными знаниями в задаче кластеризации предметных и языковых знаний.
- Разработка методов численной оценки смысловой близости ответа испытуемого варианту правильного ответа на ТЗОФ.
- Разработка архитектуры системы контроля знаний, реализующей предложенные методы и модели.

Определение 1

Конструкция ЕЯ — последовательность знаков, используемая для фиксации некоторого числа высказываний этого ЕЯ в памяти ЭВМ.

Определение 2

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

Фиксируемый СЯУ S языковой контекст представляется тройкой:

$$S = \left(O, R, T^S\right), \tag{1}$$

где O — множество объектов-участников S;
R — множество отношений между $o \in O$;
T^S — множество форм языкового описания S.

Задача 1

Дано:

- Множество ЕЯ-текстов G.

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

- Множество $V(T_i)$ ситуаций, описываемых T_i.
- Множество $M(T_i)$ объектов и/или понятий, значимых в ситуациях из множества $V(T_i)$.
- Тернарное отношение $I \subseteq G \times M \times V$, ставящее в соответствие каждому $m \in M$, $M = \bigcup_i M(T_i)$, ту ситуацию $v \in V$, $V = \bigcup_i V(T_i)$, в которой он фигурирует относительно T_i.

Далее на основе отношения I необходимо выделить группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях.

Задача 2

Дано:

- Π^R — множество правил синонимических преобразований ЕЯ-высказываний в рамках стандартных Лексических Функций (ЛФ).
- L^Π — множество пар ЕЯ-высказываний, между которыми возможно установление синонимии (относительно Π^R).
- $r(\pi)$ — условие применимости правила $\pi \in \Pi^R$.

  Для $L_i = \{T_1, T_2\} : L_i \in L^\Pi$ компонент $r(\pi)$ есть совокупность требований к $\forall w_j \in W,\ W = W_1 \cup W_2$, где $W_1 \subset T_1,\ W_2 \subset T_2$, а W_1 и W_2 — совокупности слов, заменяемых посредством π.

Требуется:

- для произвольной пары L_k ЕЯ-высказываний проанализировать условие применимости каждого правила множества Π^R и выделить образ класса $\pi \in \Pi^R$, на который объект L_k наиболее похож.

  При этом $r(\pi)$ выступает в качестве прецедента как типичного представителя таксона π.

Определение 3

Лексической Синонимической Конструкцией (ЛСК) далее именуется комплекс лексических единиц и связывающих их отношений, замена которого описывается некоторым $\pi \in \Pi^R$.

Каждой ЛСК соответствует своё ключевое слово C_0, либо непосредственно входящее в неё, либо выраженное в значениях ЛФ от C_0 в комплексе составляющих ЛСК лексических единиц.

Представим вход правила $\pi_i \in \Pi^R$ как описание заменяемого поддерева.

Тогда для всех $\pi_i \in \Pi^R$ результат анализа применимости к заданному дереву фиксируется списком пар:

$$\left\{ \left(\pi_i, C_0\left(i\right) : i = 1, \ldots, \left|\Pi^R\right| \right) \right\}, \tag{2}$$

причём в работе любого $\pi_i \in \Pi^R$ выделяются состояния: соответствующее заменяемому дереву $T_1^{\pi_i}$ и соответствующее заменяющему дереву $T_2^{\pi_i}$.

Условие $r(\pi_i)$ определяет допустимость перехода из $T_1^{\pi_i}$ в $T_2^{\pi_i}$.

Утверждение 1

Пусть R_π — множество условий применимости правила $\pi \in \Pi^R$.

В общем случае определяемый правилом π переход из T_1^π в T_2^π допустим, если $\exists r_j(\pi) \in R_\pi$: $\vee_{j=1}^m r_j(\pi) = \text{true}$, где $m = |R_\pi|$.

Обозначим $\vee_{j=1}^m r_j(\pi)$ как r_{12}. Тогда применение правила $\pi \in \Pi^R$ сводится к выполнению перехода:

$$\pi(r_{12}) : T_1^\pi \xrightarrow{\pi(r_{12})} T_2^\pi. \tag{3}$$

Отдельному правилу соответствует элементарная сеть Петри вида

$$N = \{P, T, F, H, M_0\}, \tag{4}$$

где P — множество позиций, $P = \{p_1, p_2\}$, $p_1 \Leftrightarrow T_1^\pi$, $p_2 \Leftrightarrow T_2^\pi$;

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

F и H — отображения, $F: P \times T \to \{0,1\}$, $H: T \times P \to \{0,1\}$,
для сети (4) $F(p_1, t) = 1$, $F(p_2, t) = 0$, $H(t, p_1) = 0$, $H(t, p_2) = 1$;

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

Рассмотрим $\Pi_i^R \subseteq \Pi^R$: для $\forall \pi_1 \in \Pi_i^R \;\; \exists \pi_2 \in \Pi_i^R$, $\pi_2 \neq \pi_1$, такое, что либо вход у π_2 является выходом для π_1, либо вход у π_1 есть выход у π_2.

Пусть N_i — сеть Петри, построенная из примитивов, каждый из которых есть элементарная сеть Петри и моделирует работу некоторого $\pi \in \Pi_i^R$.

Тогда последовательность применения правил $\pi \in \Pi_i^R$ моделируется последовательностью $\tau = \left(t_i^1, t_i^2, \ldots, t_i^k\right)$ срабатывания переходов сети N_i:

$$T_1^\pi \xrightarrow{\pi_1(r_{12})} T_2^\pi \xrightarrow{\pi_2(r_{23})} T_3^\pi \to \cdots \to T_k^\pi \xrightarrow{\pi_k\left(r_{k\,k+1}\right)} T_{k+1}^\pi. \tag{5}$$

При этом происходит последовательная смена разметок:

$$M_{0i} \xrightarrow{t_i^1} M_i^1 \xrightarrow{t_i^2} M_i^2 \to \cdots \to M_i^{k-1} \xrightarrow{t_i^k} M_i^k, \tag{6}$$

где $t_i^1 \Leftrightarrow \pi_1\left(r_{12}\right)$, $t_i^2 \Leftrightarrow \pi_2\left(r_{23}\right)$, $\ldots$, $t_i^k \Leftrightarrow \pi_k\left(r_{k\,k+1}\right)$,
$M_{0i} \Leftrightarrow T_1^\pi$, $M_i^1 \Leftrightarrow T_2^\pi$, $\ldots$, $M_i^{k-1} \Leftrightarrow T_k^\pi$, $M_i^k \Leftrightarrow T_{k+1}^\pi$.

Замечание

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

Пусть T_i — множество переходов сети N_i, рассматриваемое как алфавит.
Тогда задача приведения T_1^π и T_{k+1}^π к виду с одинаковой ЛСК включает:

- определение достижимости разметки M_i^k из начальной разметки M_{0i}.
  Данная задача есть поиск слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_i^k$,
  где T_i^* — множество всех слов в алфавите T_i языка $L(N_i)$;
- задача обратимости слова τ: если $\tau \in T_i^*$, то существует ли
  слово $\tau' = \left(t_i^{k'}, t_i^{(k-1)'}, \ldots, t_i^{2'}, t_i^{1'}\right)$:

$$M_{0i} \xleftarrow{t_i^{1'}} M_i^1 \xleftarrow{t_i^{2'}} M_i^2 \leftarrow \cdots \leftarrow M_i^{k-1} \xleftarrow{t_i^{k'}} M_i^k, \qquad (7)$$

  где $M_{0i} \Leftrightarrow T_1^\pi,\ M_i^1 \Leftrightarrow T_2^\pi,\ \ldots,\ M_i^{k-1} \Leftrightarrow T_k^\pi,\ M_i^k \Leftrightarrow T_{k+1}^\pi$;
- задача определения оптимального слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_i^k$.
  Суть: если существуют $\tau_1, \tau_2, \ldots, \tau_l$, описывающих смену разметок

$$M_{0i} \xrightarrow{\tau_1} M_i^k,\ M_{0i} \xrightarrow{\tau_2} M_i^k,\ \ldots,\ M_{0i} \xrightarrow{\tau_l} M_i^k,$$

  то результат есть обратимое слово минимальной длины.

Лемма 1

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

Теорема 1

Сеть N_i безопасна в течение всего времени функционирования системы.

Теорема 2

Все символы-переходы $t_i^j \in T_i$ сети N_i различны.

Теорема 3

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

Теорема 4

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

Моделирование входа/выхода правила Δ-грамматики

Рассмотрим сеть Петри:

$$N_{\pi(k)} = \left\{P_{\pi(k)}, T_{\pi(k)}, F_{\pi(k)}, H_{\pi(k)}, C, M_{0\pi(k)}\right\}, \tag{8}$$

где $p^i_{\pi(k)} \in P_{\pi(k)}$ отождествляется с пройденным узлом w^π_i дерева T^π_k;

отдельному $t^i_{\pi(k)} \in T_{\pi(k)}$ соответствует совокупность требований лексической, грамматической части и метки входящей ветви узла w^π_i;

$F_{\pi(k)}$ и $H_{\pi(k)}$ — матрицы инцидентности;

$C = \{c_1, c_2, c_3, c_4, c_5\}$ — множество цветов маркера:

- c_1 — анализ применимости правила;
- c_2 — синтез дерева на выходе правила;
- c_3 — определение ключевого слова ЛСК;
- c_4 — расстановка композиционных меток;
- c_5 — запрет срабатывания перехода;

$M_{0\pi(k)}$ — начальная разметка.

Теорема 5

Все порождаемые сетью $N_{\pi(k)}$ процессы конечны.

Теорема 6

Сеть $N_{\pi(k)}$ является ограниченной.

Определение 4

Помеченные деревья T_1 и T_1' изоморфны с точностью до функционального соответствия, если в дереве T_1' из узла α_{11}' в α_{12}' идёт ветвь с некоторой пометкой только тогда, когда в дереве T_1 из α_{11} в α_{12} идёт ветвь с той же пометкой. При этом узел α_{11}' отвечает требованиям, содержащимся в узле α_{11}, а узел α_{12}' — требованиям узла α_{12}. В таком случае считается, что узел α_{11}' функционально соответствует узлу α_{11}, а узел α_{12}' — узлу α_{12}.

Пусть $T_{\chi 1} = \langle W_{\chi 1}, V_{\chi 1} \rangle$ и $T_{\chi 2} = \langle W_{\chi 2}, V_{\chi 2} \rangle$ — помеченные деревья, где $W_{\chi 1}$ и $W_{\chi 2}$ — множества узлов, $V_{\chi 1}$ и $V_{\chi 2}$ — множества ветвей.

Теорема 7

Задача установления функционального соответствия деревьев $T_{\chi 1}$ и $T_{\chi 2}$ принадлежит классу P комбинаторных задач с временной оценкой n^D, где $n = \max(|W_{\chi 1}|, |W_{\chi 2}|,)$, $D = \sum_{i=1}^{|V^R|} \varphi(a_i)$, φ есть матрица ограничений на характер ветвления, V^R — словарь пометок на ветвях.

Пусть слово w_i, заменяется некоторым $\pi \in \Pi^R$. Опишем Лексическое Значение (ЛЗ) слова w_i посредством структуры:

$$Lm\left(w_i\right) = \left(w_i, L^M\right), \tag{9}$$

где элемент списка L^M может представлять как бинарное отношение R_2 между парой понятий C_1 и C_2:

$$M_p = \left(R_2, C_1, C_2\right), \tag{10}$$

так и рекурсивно определяемые отношения произвольной арности:

$$M'_p = \left(R_n, C, L^M\right) \text{ и} \tag{11}$$

$$M''_p = \left(R_c, L^M\right), \tag{12}$$

где $R_c \in \{\vee, \&, \neg\}$. Посредством L^M в (11) задается связь понятия C с другими словами и понятиями.

Утверждение 2

Если имеется описание лексического значения слова w_i посредством структуры $Lm\left(w_i\right) = \left(w_i, L^M\right)$, то смысл слова определяется набором характеристических функций ChF_{hi} таких, что выполняются условия:

1. в списке L^M содержится структура $M_p = (R_2, C_1, C_2)$ вида (10) (обозначим её как ChF_{Val}), при этом $ChF_{hi}\left(w_i\right) = C_2$, где C_2 — некоторое известное понятие, а L^M может быть третьим аргументом структуры (11);
2. существует структура (далее обозначаемая как ChF_{Name}) либо вида (10) и при этом $M_p = \left(ChF_{hi}, C_1^1, C_2^1\right)$, либо вида (11) и $M_p' = \left(ChF_{hi}, C, L^M\right)$, но в обоих случаях ChF_{hi} — имя известного смыслового отношения;
3. если ChF_{Name} есть первая структура, удовлетворяющая условию (2) при обратном просмотре списка L^M от ChF_{Val}, и $L^{M'} \subset L^M$ есть список такой, что либо $L^{M'} = \left\{\left(ChF_{hi}, C_1^1, C_2^1\right), \ldots, ChF_{Val}\right\}$, либо $L^{M'} = \left\{\left(ChF_{hi}, C, L^M\right) : ChF_{Val} \in L^M\right\}$, то каждое следующее утверждение в $L^{M'}$ имеет общий аргумент-обозначение переменной с предыдущим утверждением.

Введём в рассмотрение многозначный формальный контекст:

$$K^{LM} = \left(G^{LM}, M^{LM}, V^{LM}, I^{LM}\right), \tag{13}$$

где $\forall g^{LM} \in G^{LM}$ есть вариант толкования ЛЗ слова w_i, $g^{LM} = Lm_j\left(w_i\right)$;

множество признаков $M^{LM} = M_1^{LM} \cup M_2^{LM}$, при этом

если $m^{LM} \in M_1^{LM}$ то $m^{LM} = ChF_{hi}\left(w_i\right)$,

если $m^{LM} \in M_2^{LM}$ то существует структура (R, C_1, C_2) вида (10),

причём $m^{LM} = R$, R — известное отношение, $\left(R, C_1^1, C_2^1\right) \in L^{M'}$,

$L^{M'}$ формируется согласно *условию (3) утверждения 2* .

множество признаковых значений $V^{LM} = V_1^{LM} \cup V_2^{LM}$, при этом

если $v^{LM} \in V_1^{LM}$ то v^{LM} — имя характеристической функции ChF_{hi}, для которой определено значение $ChF_{hi}\left(w_i\right)$,

если $v^{LM} \in V_2^{LM}$ то $v^{LM} = ChF'_{hi}\left(w'_i\right)$: $Lm\left(w'_i\right) = \left(w'_i, L^{M'}\right)$,

где $L^{M'}$ сформирован согласно *условию (3) утверждения 2* ;

отношение $I^{LM} \subseteq G^{LM} \times M^{LM} \times V^{LM}$ ставит в соответствие каждой характеристической функции её значение для заданного w_i.

Определение 5

Формальное Понятие (ФП) для формального контекста (13) есть пара

$$\left(X^{LM}, Y^{LM}\right) : X^{LM} \subseteq G^{LM}, Y^{LM} \subseteq M^{LM} \times V^{LM},$$

$$X^{LM} = Y^{LM'}, Y^{LM} = X^{LM'}, \text{причем}$$

$$X^{LM'} = \left\{ \left(m^{LM}, v^{LM}\right) : m^{LM} \in M^{LM}, \right.$$

$$\left. v^{LM} \in V^{LM} \;\middle|\; \forall g^{LM} \in X^{LM} : m^{LM}\left(g^{LM}\right) = v^{LM} \right\},$$

$$Y^{LM'} = \left\{ g^{LM} \in G^{LM} \;\middle|\; \forall \left(m^{LM}, v^{LM}\right) \in Y^{LM} : m^{LM}\left(g^{LM}\right) = v^{LM} \right\}.$$

Определение 6

ФП $\left(X_1^{LM}, Y_1^{LM}\right)$ является подпонятием для ФП $\left(X_2^{LM}, Y_2^{LM}\right)$, если $X_1^{LM} \subseteq X_2^{LM}$, а $Y_2^{LM} \subseteq Y_1^{LM}$: $\left(X_1^{LM}, Y_1^{LM}\right) \leq \left(X_2^{LM}, Y_2^{LM}\right)$. При этом ФП $\left(X_2^{LM}, Y_2^{LM}\right)$ называют суперпонятием для ФП $\left(X_1^{LM}, Y_1^{LM}\right)$, а отношение $\leq$ — отношением порядка для формальных понятий.

Определение 7

Множество всех формальных понятий заданного формального контекста вместе с отношением порядка называют решёткой формальных понятий, далее для обозначения решёток ФП будем использовать символ $\Re$.

Определение 8

Пусть $C^{\Re} \subset \Re^{LM}\left(G^{LM}, M^{LM}, V^{LM}, I^{LM}\right)$. Формальное понятие $\left(X^{LM}, Y^{LM}\right)$ называется Наименьшим Общим Суперпонятием (НОСП) для $C^{\Re}$, если $\left(X_i^{LM}, Y_i^{LM}\right) \leq \left(X^{LM}, Y^{LM}\right)$ для $\forall \left(X_i^{LM}, Y_i^{LM}\right) \in C^{\Re}$ и $\nexists (X, Y) \in \Re^{LM}\left(G^{LM}, M^{LM}, V^{LM}, I^{LM}\right) \setminus C^{\Re}$: $(X, Y) \leq \left(X^{LM}, Y^{LM}\right)$ и $\left(X_i^{LM}, Y_i^{LM}\right) \leq (X, Y)$ для $\forall \left(X_i^{LM}, Y_i^{LM}\right) \in C^{\Re}$. Наибольшее Общее Подпонятие (НОПП) для $C^{\Re}$ определяется аналогично.

Определение 9

Под областью в решётке формальных понятий понимается набор ФП, связанных отношением $\leq$ с одним НОПП и/или одним НОСП.

Утверждение 3

Утверждения $\left(R_n, C, L_1^M\right)$ и $\left(R_n, C, L_2^M\right)$ вида (11) представимы одним «ИЛИ»-утверждением

$$\left(R_n, C, \left\{\left(\vee, L_3^M\right)\right\}\right),$$

если наборы ФП, полученные на основе L_1^M, L_2^M и L_3^M, образуют области $\Re^{LM}\left(G_1^{LM}, M_1^{LM}, V_1^{LM}, I^{LM}\right)$, $\Re^{LM}\left(G_2^{LM}, M_2^{LM}, V_1^{LM}, I^{LM}\right)$ и, соответственно, $\Re^{LM}\left(G_3^{LM}, M_3^{LM}, V_1^{LM}, I^{LM}\right)$ с НОСП, которое имеет R_n в качестве значения признака. При этом

$$G_1^{LM} = \left\{\left(w_i^1, L_1^M\right)\right\}, G_2^{LM} = \left\{\left(w_i^2, L_2^M\right)\right\},$$

$$M_1^{LM} \neq M_2^{LM}, M_3^{LM} = M_1^{LM} \cup M_2^{LM},$$

$$\Re^{LM}\left(G_3^{LM}, M_3^{LM}, V_1^{LM}, I^{LM}\right) =$$

$$= \Re^{LM}\left(G_1^{LM}, M_1^{LM}, V_1^{LM}, I^{LM}\right) \cup \Re^{LM}\left(G_2^{LM}, M_2^{LM}, V_1^{LM}, I^{LM}\right).$$

Утверждение 4

Утверждения $\left(R_n, C, L_1^M\right)$ и $\left(R_n, C, L_2^M\right)$ вида (11) представимы одним «И»-утверждением

$$\left(R_n, C, \left\{\left(\wedge, L_3^M\right)\right\}\right),$$

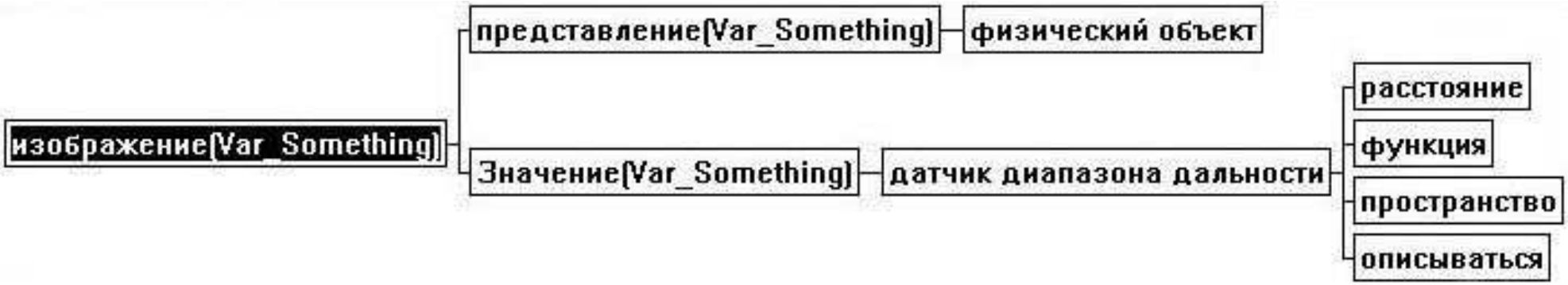
если на основе L_1^M, L_2^M и L_3^M определяются формальные понятия

$$(X, Y_1), (X, Y_2) \text{ и } (X, Y_3) : Y_3 = Y_1 \cup Y_2.$$

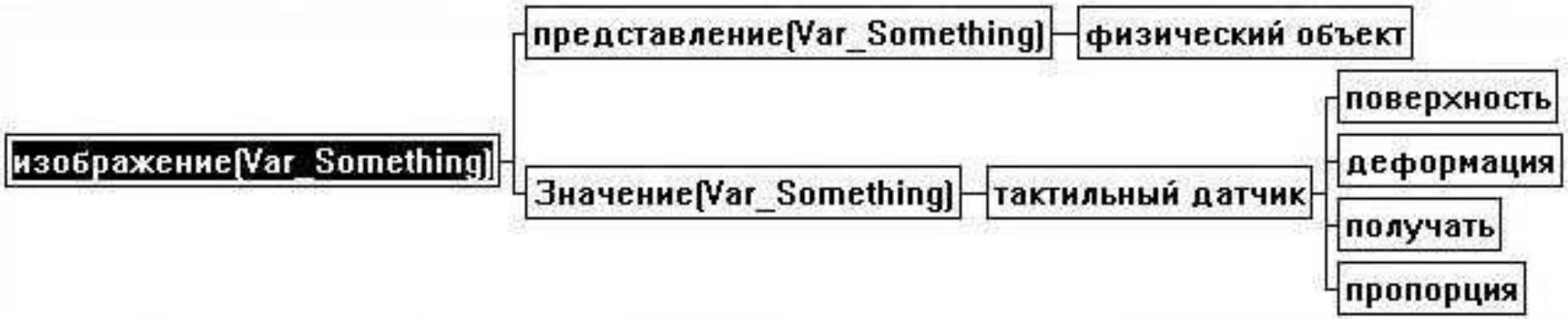
Замечание

Согласно *утверждению 2*, внешне различные описания структур вида (9) для одного и того же лексического значения задают единое множество характеристических функций. Следовательно, мощность указанного множества не зависит от числа обобщаемых структур.

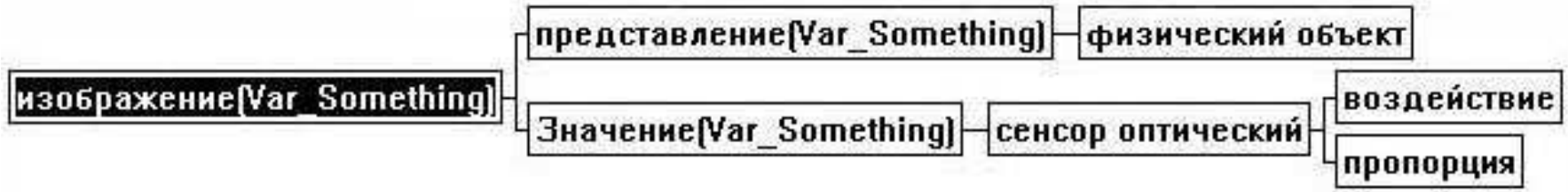
Толкование 1 и характеристические функции ЛЗ слова «изображение»



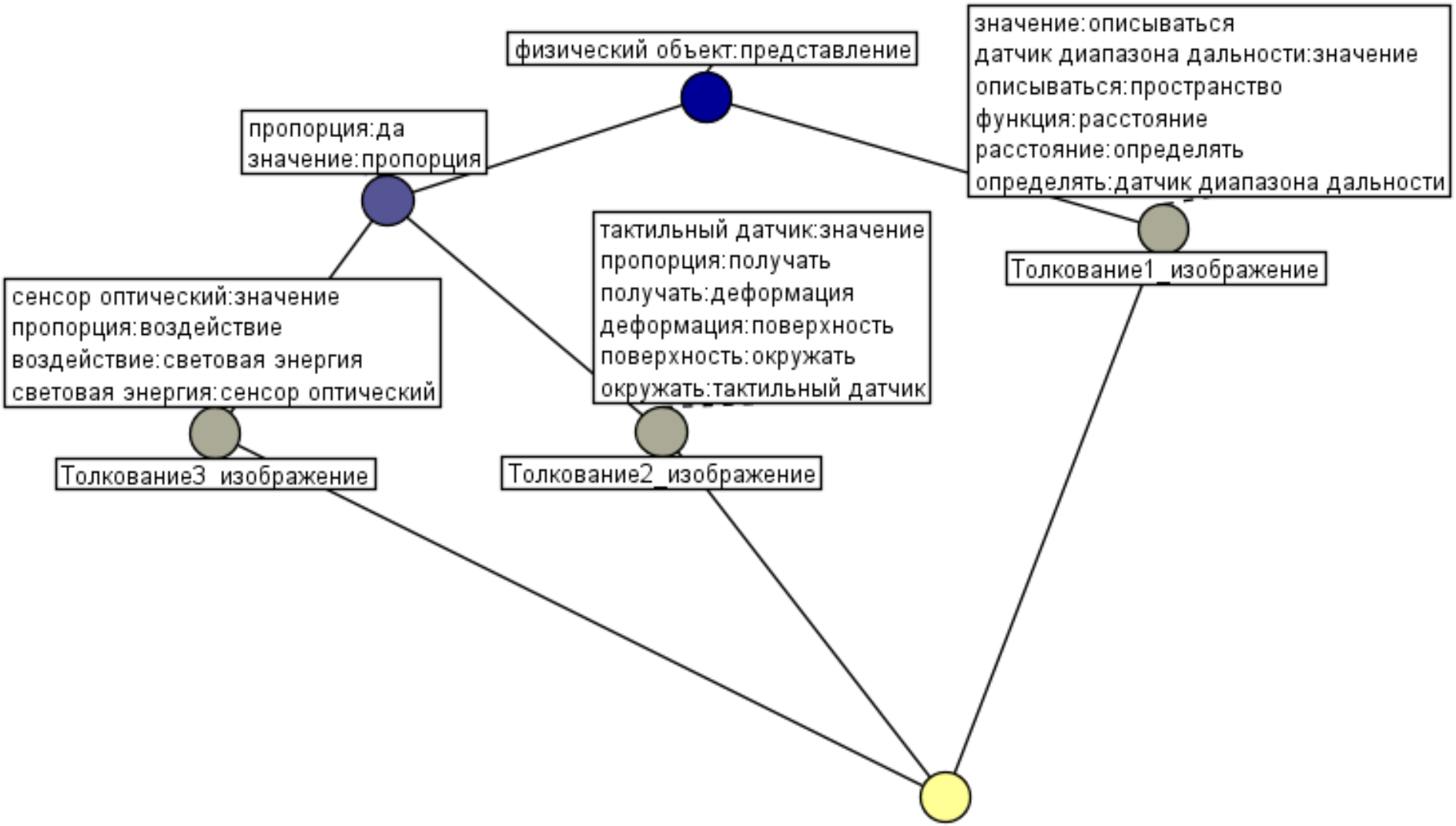
Толкование 2 и характеристические функции ЛЗ слова «изображение»



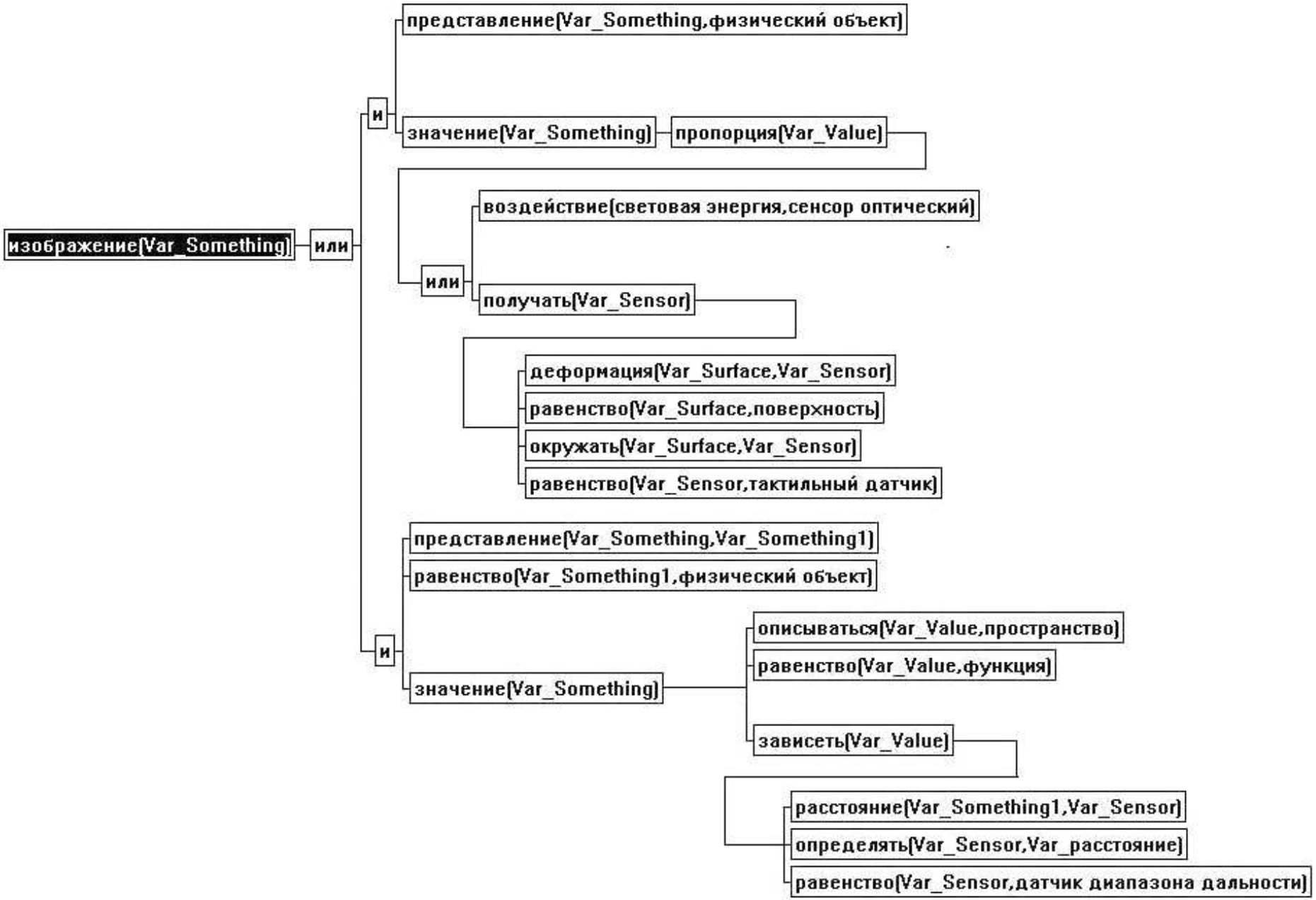
Толкование 3 и характеристические функции ЛЗ слова «изображение»



Решетка формальных понятий для независимых толкований



Обобщение независимых толкований ЛЗ слова «изображение»



Пусть T^S — множество СЭ-фраз, определяющих некоторую СЯУ S.

При рассмотрении $\forall T_i \in T^S$ как множества символов справедливо:

$$T_i = T_i^C \cup T_i^F,$$

где T_i^C — общая неизменная часть всех $T_i \in T^S$, T_i^F — изменяемая часть.

Пусть W_{ij} — буквенный состав слова, j — его порядковый номер во фразе.

Тогда

$$W_{ij} = W_{ij}^C \cup W_{ij}^F,$$

где $W_{ij}^C \subset T_i^C$ — неизменная, $W_{ij}^F \subset T_i^F$ — флективная часть.

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

- W_{ij}^C и W_{ij}^F каждого W_{ij} при $\left|W_{ij}^C\right| \to \max$;
- синтаксическое отношение R_q, определяющее допустимость сочетания слов с буквенным составом флексий W_{ij}^F и W_{ik}^F, $k \neq j$.

Пусть J — индексное множество для неизменных частей всех слов, употребленных во всех СЭ-фразах множества T^S.

Определение 10

Моделью L линейной структуры фразы $T_i \in T^S$ назовем упорядоченную совокупность индексов $j \in J$ неизменных частей слов, входящих в T_i.

Пусть $h\left(j, L\left(T_i\right)\right)$ — позиция индекса j в модели $L\left(T_i\right)$, где $j \in J$.

Тогда множество связей относительно $L\left(T_i\right)$

$$D : T_i \to \left\{ \left(h\big(j, L(T_i)\big),\ h\big(k, L(T_i)\big) \right) :\ j \neq k \right\}.$$

Определение 11

Связь $d_{qi} = \left(h\big(j, L\left(T_i\right)\big),\ h\big(k, L\left(T_i\right)\big) \right)$ допустима для модели $L\left(T_i\right)$, если существует пара СЭ-фраз $\{T_l, T_m\} \subset T^S$ таких, что и $L\left(T_l\right)$, и $L\left(T_m\right)$ имеют подпоследовательностью либо $\{j, k\}$, либо $\{k, j\}$.

Пусть для любого $T_i \in T^S$ все $d_{qi} \in D(T_i)$ удовлетворяют *определению 11*.

Определение 12

Будем считать, что модель $L(T_i)$ проективна относительно множества синтаксических связей в T^S, если

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

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

Замечание

Сосуществование словоформ в линейном ряду относительно $L(T_i)$ определяется синтагматическими зависимостями, которые выражаются на множестве T_i^F и задаются отношениями из множества R в составе структуры (1). Для построения множества R нужно найти совокупность моделей линейных структур фраз из T^S, отвечающих *определению 12*.

Пусть $\bigcup_i D(T_i)$ есть множество связей, допустимых для $\forall L(T_i)$: $T_i \in T^S$.

Определение 13

Множество пар (j, k), сгруппированных по некоторому общему для них 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 $j \neq m$.

Пусть $G^F = \left\{ f_{ij} : f_{ij} = \odot \left(W_{ij}^F \right) \right\}$, $I^F = \left\{ (f_{ij}, f_{ik}) : s\,(j, k) = \text{true} \right\}$,

где $\odot$ — последовательная конкатенация символов.

Отношение s задается рекурсивно на основе $\left(V^J, I^J\right)$ следующим образом.

1. $s\,(j_1, j_1) = \text{true}$.
2. $s\,(j_1, j_2) = \text{true}$, если выполняется одно из двух условий:
   - $\exists\, E_1 \in V^J$: $(j_1, j_2) \in E_1$, причем $\exists\, j_3 \in J : s\,(j_2, j_3) = \text{true}$;
   - $\exists\, (E_1, E_2) \in I^J : \exists\, j_3 \in J$: $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$, $s\,(j_3, j_2) = \text{true}$.

Отношению I^F соответствует формальный контекст *сочетаемости флексий*:

$$K^F = \left(G^F, M^F, I^F \right), \text{ в котором } M^F = G^F.$$

Пусть $W_{ij} \subset T_i$, где $T_i \in T^S$. Рассмотрим $T_i^{\odot} = \left\{w_{ij} : w_{ij} = \odot\left(W_{ij}\right)\right\}$.

Положим также, что $\exists\, T_i^P \subset T_i$, определяющее последовательность

$$P_i^{\odot} = \left\{u_k : u_k = \odot\left(W_k^P\right), \bigcup_k W_k^P = T_i^P\right\}.$$

Лемма 2

Последовательность $P_i^{\odot}$ содержит слово-предикат, если $\exists\, \{j, 0, k\} \subset L\left(T_i\right)$: $\{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{\odot}$, где $\{u_1, \ldots, u_p\} = P_i^{\odot}$, а $p = \left|P_i^{\odot}\right|$.

Лемма 3

Слово $u_k \in P_i^{\odot}$ входит в состав Расщепленного Предикатного Значения (РПЗ), если $\exists\, T_j \in T^S : L\left(T_j\right) \neq L\left(T_i\right)$, а $u_k \in P_j^{\odot}$. При этом $\nexists\, T_k \in T^S$, для которого $P_k^{\odot} \subset P_i^{\odot}$, а $L\left(T_k\right) \neq L\left(T_j\right)$ и $L\left(T_k\right) \neq L\left(T_i\right)$.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих *лемме 3*.

Теорема 8

Для построения формального контекста K^F при наличии РПЗ необходимо и достаточно найти множество $T' \subset T^S$: $T' = \left\{T_i : \left|P_i^{\odot\prime}\right| \to \max\right\}$.

Пусть $\left(V^J, I^J\right)$ — граф синтагм, J — индексное множество, на котором задаются $L\left(T_i\right) : T_i \in T^S$. Рассмотрим

$$I_1^J = \left\{(j,k) \colon \exists\, E \in V^J, (j,k) \in E\right\}.$$

Назовем $\left(V_1^J, I_1^J\right)$, $V_1^J = J$, деревом-прецедентом для T^S.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих *лемме 3*, а

$T' \subset T^S$ — множество, рассматриваемое *теоремой 8*.

Для $u_k \in \bigcup_i P_i^{\odot\prime} \colon T_i \in T'$ неизменная и флективная части формируются сравнением буквенного состава со всеми $u_j \in \bigcup_l P_l^{\odot} \colon T_l \in \left(T^S \setminus T'\right)$.

При этом необходимо, чтобы $2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|$, где индексы C относятся к составу неизменной, а F — флективной части слова.

Дерево $\left(V_1^J, I_1^J\right)$ преобразуется следующим образом:

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

Пусть разбором некоторого $T_i \in T^S$ выделена последовательность

$$S_{ki} = \left\{ v_1, \ldots, v_{n(k,i)}, m_{ki} \right\}. \tag{15}$$

где v_1 — предикатное слово (глагол или отглагольное существительное);

k — порядковый номер последовательности среди выявленных из T_i;

$n(k,i)$ — число соподчиненных существительных $\left\{ v_2, \ldots, v_{n(k,i)}, m_{ki} \right\}$.

Утверждение 5

При одновременном наличии $S_{ki} = \left\{ v_1, \ldots, v_{n(k,i)}, m_{ki} \right\}$ и $S_{1ki} = \{ v_1, m_{ki} \}$ в разных текстах множества T^S структуры (1) для заданной СЯУ имеет место частичная смысловая эквивалентность (относительно m_{ki}).

Утверждение 6

При $R_q(v_1, v_2) = \text{true}$ возможно установление указанного отношения между v_1 и любым словом последовательности (15).

Замечание

Как следует из транзитивности R_q, $\forall\, v_l \in \left\{ v_2, \ldots, v_{n(k,i)}, m_{ki} \right\}$ в составе последовательности (15) обозначает понятие, значимое в ситуации v_1.

Пусть L^S — множество моделей линейных структур ЕЯ-фраз из T^S на J.

Лемма 4

Пара $\{j_1, j_2\} \subset J$ соответствует синонимам, если $\exists \{L(T_1), L(T_2)\} \subseteq L^S$: $L(T_1) = J_1 \odot \{j_1\} \odot J_2$ и $L(T_2) = J_1 \odot \{j_2\} \odot J_2$, где $J_1 \subset J$, $J_2 \subset J$.

Пусть P^J — множество пар, отвечающих *лемме 4*, а $L^{S'}$ формируется из L^S заменой индексов, вошедших в пары из P^J, на некоторые $j \in (\mathbb{N} \backslash J)$.

Теорема 9

Индексы с максимальной встречаемостью относительно $L^{S'}$ соответствуют существительным, непосредственно подчиненным предикатному слову.

Пусть J^N — множество индексов, отвечающих *теореме 9*, а множество моделей линейных структур фраз $T_i \in T^S$ на множестве J^N — как L^N.

Положим, $\exists\, L_j^{S'} \subset L^{S'}$: для $\forall L_1(T_i) \in L_j^{S'}$ все $L_2(T_i) \in L^N$ одинаковы и соответствуют некоторой $L_2(T_j) \in L^N$, $T_j \in T^S$.

Теорема 10

Индексы $j \notin J^N$ с максимальной встречаемостью в разных $L_1(T_i) \in L_j^{S'}$ соответствуют либо наречиям, либо прилагательным, либо опорным существительным в составе генитивных конструкций.

Обозначим множество индексов, удовлетворяющих *теореме 10*, как J^A.

Синтаксическе роли и флексии для слов с индексами из

$$\left(\left(J \setminus J^P\right) \cup J^{P'}\right) \setminus \left(J^N \cup J^A\right) \cup \{0\}$$

выделяются аналогично случаю РПЗ, рассматриваемого *теоремой 8*.

При этом вместо индексов с ненулевым значением рассматриваются индексы из $J^N \cup J^A$.

Представим языковой контекст СЯУ посредством формального контекста:

$$K^S = \left(G^S, M^S, I^S\right), \tag{16}$$

где $\forall g \in G^S$ — основа слова, синтаксически подчинённого другому слову из некоторой $T_i \in T^S$ в составе структуры (1).

Множество признаков M^S включает подмножества, содержащие:

- указания на основу синтаксически главного слова (M_1);
- указания на флексию главного слова (M_2);
- связи «основа–флексия» для синтаксически главного слова (M_3);
- сочетания флексий зависимого и главного слова (M_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 (M_5).

Посредством $I^S \subseteq G^S \times M^S$ выделяются классы отношений из R в (1) по сходству основы главного, флексии зависимого слова, лексической и флективной сочетаемости.

Теорема 11

Пусть $\{m_1, m_2, m_3\} \subset M_1^S$. Если считать признаки m_1, m_2 и m_3 взаимно различными, то m_1 соответствует указанию на основу главного слова, m_2 — зависимого слова РПЗ, m_3 — однословного смыслового эквивалента этого РПЗ при выполнении трех условий:

- $\exists\, g_1 \in G^S$: $I^S(g_1, m_1) = \text{true}$, $I^S(g_1, m_3) = \text{false}$, $m_2 = p_{bs} \odot g_1$. Здесь p_{bs} есть обозначение символьной константы «главное-основа:»;
- $\exists\, \{g_2, g_3\} \subset G$, при этом объекты g_1, g_2 и g_3 взаимно различны, а

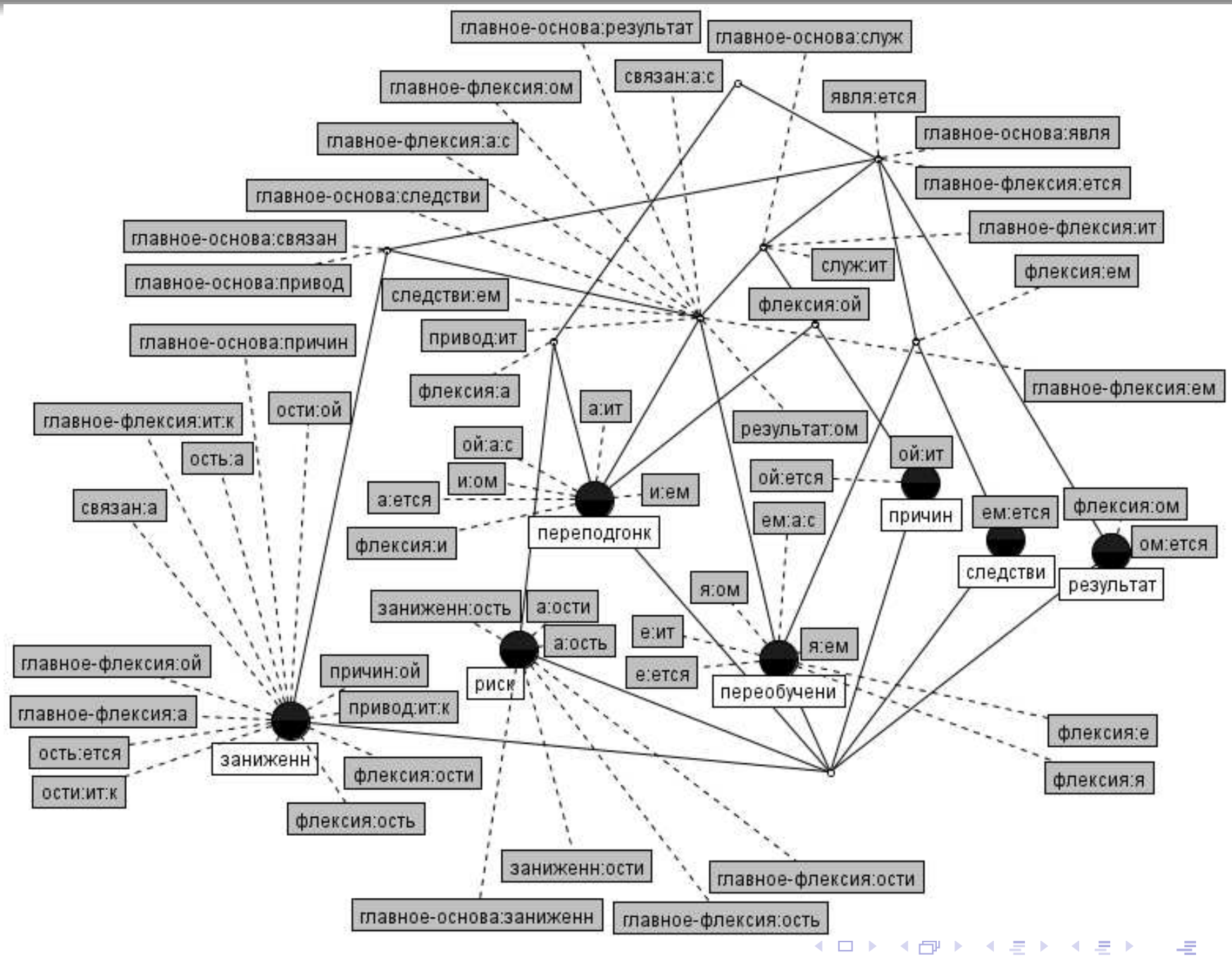
$$I^S(g_2, m_3) \wedge I^S(g_3, m_3) \wedge \\ \wedge \big(I^S(g_2, m_1) \wedge I^S(g_3, m_2) \vee \\ \vee I^S(g_2, m_2) \wedge I^S(g_3, m_1) \big) = \text{true};$$

- не существует других троек объектов, для которых признак m_3 занимал бы место либо m_1, либо m_2 в вышеуказанных соотношениях.

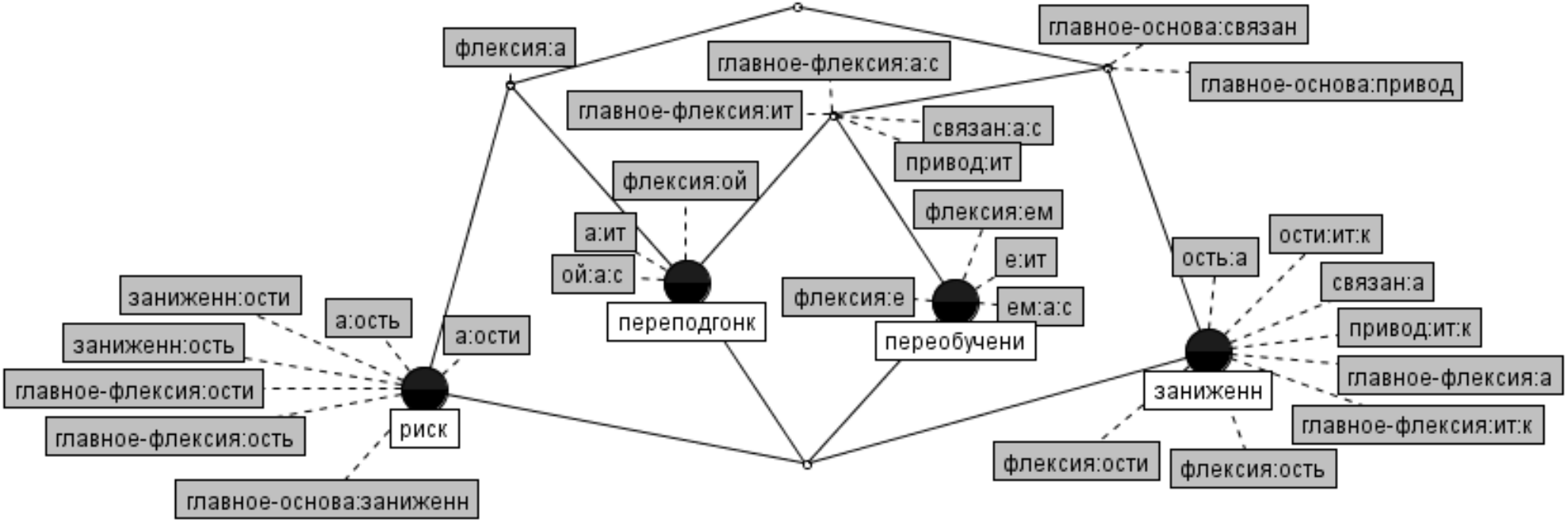
Замечание

После удаления информации РПЗ формальный контекст (16) отражает классы отношений для ролей объектов-участников ситуации.

Пример формального контекста СЯУ до удаления информации РПЗ



Формальный контекст после удаления информации РПЗ



Рассмотрим модель тезауруса в виде формального контекста:

$$K^{TH} = \left(G^{TH}, M^{TH}, I^{TH}\right), \tag{17}$$

где G^{TH} состоит из символьных пометок отдельных СЯУ.

M^{TH} содержит признаки формальных контекстов всех $g^{TH} \in G^{TH}$.

Кроме того, в составе M^{TH} выделяются подмножества:

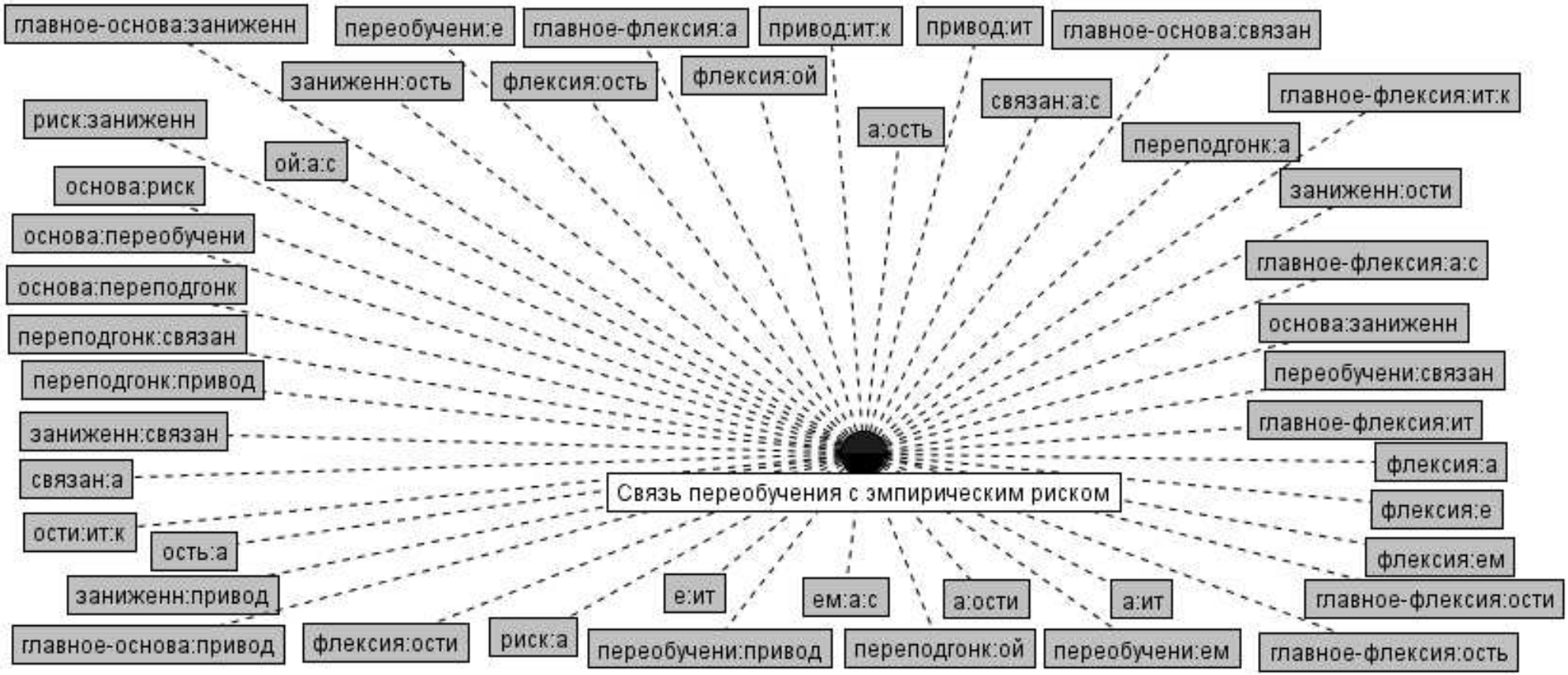
- M_6 — указаний на объекты формальных контекстов вида (16) отдельных $g^{TH} \in G^{TH}$;
- M_7 — множество связей «основа–флексия» для синтаксически зависимого слова;
- M_8 — множество сочетаний основ зависимого и главного слова.

Пусть $K^E = \left(G^E, M^E, I^E\right)$ есть формальный контекст СЯУ S_1 корректного описания некоторого факта, $K^X = \left(G^X, M^X, I^X\right)$ — формальный контекст произвольной СЯУ S_2, а $M^U = M_6 \cup M_7 \cup M_8 \cup M_4^E \cup M_4^X \cup M_5^E \cup M_5^X$.

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

Простейший случай *схожести* S_1 и S_2: для $\forall g^X \in G^X \; \exists\, g^E \in G^E : g^X = g^E$ и любой признак $m^E \in M^E$ объекта g^E относится к g^X.

Пример объекта отдельной СЯУ в формальном контексте тезауруса



Случай 2 соответствия объектов

$g^X = g^E$, условие *простейшего случая* не выполняется, но существует объект $g^{TH} \in G^{TH}$, обладающий признаком $m_1^{TH} \in M_6 : m_1^{TH} = p_b \odot g^E$ при обязательном выполнении следующих условий:

$$\left(\exists\, m_{fl}^E \in M_5^E : m_{fl}^E = p_{fl} \odot f^E\right) \rightarrow \left(\exists\, m_{17}^{TH} \in M_7 : m_{17}^{TH} = g^E \odot «:» \odot f^E\right),$$

$$\text{при этом } \left(I^E\left(g^E, m_{fl}^E\right) \wedge I^X\left(g^E, m_{fl}^E\right)\right) \rightarrow I^{TH}\left(g^{TH}, m_{17}^{TH}\right);$$

$$\left(\exists\, m_{bs}^E \in M_1^E : m_{bs}^E = p_{bs} \odot b^E\right) \rightarrow \left(\exists\, m_{18}^{TH} \in M_8 : m_{18}^{TH} = g^E \odot «:» \odot b^E\right),$$

$$\text{при этом } I^E\left(g^E, m_{bs}^E\right) \rightarrow I^{TH}\left(g^{TH}, m_{18}^{TH}\right);$$

$$\left(\exists\, m_{bs}^X \in M_1^X : m_{bs}^X = p_{bs} \odot b^X\right) \rightarrow \left(\exists\, m_{28}^{TH} \in M_8 : m_{28}^{TH} = g^E \odot «:» \odot b^X\right),$$

$$\text{при этом } I^X\left(g^E, m_{bs}^X\right) \rightarrow I^{TH}\left(g^{TH}, m_{28}^{TH}\right).$$

Кроме того, для $\forall\, m^{TH} \in \left(M^{TH} \setminus M^U\right)$ верно:

$$I^{TH}\left(g^{TH}, m^{TH}\right) \rightarrow \left(I^E\left(g^E, m^{TH}\right) \wedge I^X\left(g^E, m^{TH}\right)\right). \qquad (18)$$

Случай 3 соответствия объектов

$g^X \neq g^E$, но существует объект $g^{TH} \in G^{TH}$, обладающий признаками

$$m_1^{TH} \in M_6 : m_1^{TH} = p_b \odot g^E \text{ и}$$
$$m_2^{TH} \in M_6 : m_2^{TH} = p_b \odot g^X,$$

при этом для любого признака $m^{TH} \in \left(M^{TH} \backslash M^U\right)$ справедливо:

$$I^{TH}\left(g^{TH}, m^{TH}\right) \rightarrow \left(I^E\left(g^E, m^{TH}\right) \wedge I^X\left(g^X, m^{TH}\right)\right). \quad (19)$$

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 (17) на наборе признаков с префиксом p_{bs} для одной и той же СЯУ.

Случай 4 соответствия объектов

$g^X \neq g^E$, но существует $g_1^{TH} \in G^{TH}$, обладающий признаком $m_1^{TH} \in M_6$: $m_1^{TH} = p_b \odot g^E$, а для $\forall\, m^E \in \left(M_4^E \cup M_5^E\right)$ верно то, что

$$\left(I^{TH}\left(g_1^{TH}, m_1^{TH}\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^{TH}\left(g_1^{TH}, m^E\right).$$

При этом существуют признаки $m_2^{TH} \in M_6$ и $m^X \in \left(M_1^X \cup M_2^X \cup M_3^X\right)$:

$$\left(I^{TH}\left(g_1^{TH}, m_2^{TH}\right) \wedge I^X\left(g^X, m^X\right)\right) \rightarrow I^{TH}\left(g_1^{TH}, m^X\right),$$

где $m_2^{TH} = p_b \odot g^{X_1}$, $g^{X_1} \neq g^X$, а пара $\left(g^{X_1}, g^E\right)$ соответствует *Случаю 3* при генерации формального контекста для g_1^{TH}.

В то же время существует объект $g_2^{TH} \in G^{TH}$, относительно которого пара $\left(g^X, g^{X_1}\right)$ также будет соответствовать *Случаю 3*.

Генерируемый при этом формальный контекст для g_2^H обозначим далее как K^{X_1}. По аналогии с K^E и K^X, $K^{X_1} = \left(G^{X_1}, M^{X_1}, I^{X_1}\right)$.

Оценка схожести ситуаций языкового употребления S_1 и S_2 относительно их формальных контекстов $K^E = (G^E, M^E, I^E)$ и $K^X = (G^X, M^X, I^X)$, из которых удалена информация РПЗ , вычисляется по формуле:

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{20}$$

где $n = |G^X|$,

spc_k есть численное значение схожести объектов в паре (g_k^X, g^E).

Если (g_k^X, g^E) 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 g^E и g^X значение spc_k равно 1.0.

Если пара $\left(g_k^X, g^E\right)$ отвечает одному из *Случаев 2–3* соответствия объектов, то оценка схожести g_k^X и g^E вычисляется по формуле:

$$-\log_2\left(1-\frac{D_c}{path_C}\right)\times\frac{\left|B^{LCS}\right|}{\left|B_1\backslash B^{LCS}\right|+\left|B_2\backslash B^{LCS}\right|+\left|B^{LCS}\right|}, \qquad (21)$$

где $D_c = 2$, число $path_C = 4$.

В множество B^{LCS} войдут признаки $m^{TH} \in \left(M^{TH}\backslash M^U\right)$, для которых справедливо либо соотношение (18) для *Случая 2*,
либо соотношение (19) для *Случая 3*.

При этом

$$B_1 = \left\{ m^E : m^E \in \left(M_1^E \cup M_2^E \cup M_3^E \right), I^E\left(g^E, m^E\right) = \text{true} \right\},$$
$$B_2 = \left\{ m^X : m^X \in \left(M_1^X \cup M_2^X \cup M_3^X \right), I^X\left(g_k^X, m^X\right) = \text{true} \right\}.$$

Если пара $\left(g_k^X, g^E\right)$ отвечает *Случаю 4* соответствия объектов, то оценка схожести g_k^X и g^E вычисляется по формуле:

$$-\log_2\left(1-\frac{D_c}{path_C}\right)\times\frac{\left|B^{LCS}\right|}{\left|B_1\backslash B^{LCS}\right|+\left|B_2\backslash B^{LCS}\right|+\left|B^{LCS}\right|}. \qquad (22)$$

Для рассматриваемого случая имеем:

$$B_1=\left\{\, m^{X_1} : m^{X_1}\in\left(\, M_1^{X_1}\cup M_2^{X_1}\cup M_3^{X_1}\,\right), I^{X_1}\left(g^{X_1}, m^{X_1}\right)=\text{true}\,\right\},$$
$$B_2=\left\{\, m^{X} : m^{X}\in\left(\, M_1^{X_1}\cup M_2^{X_1}\cup M_3^{X_1}\,\right), I^{X_1}\left(g_k^{X}, m^{X}\right)=\text{true}\,\right\},$$

где $D_c=2$, $\left(M_1^{X_1}\cup M_2^{X_1}\cup M_3^{X_1}\right)\subset M^{X_1}$, $B^{LCS}=B_1\cap B_2$.

Соответствие *Случаю 4* обычно проверяется в несколько итераций.

В ходе каждой последующей итерации число признаков, не являющихся общими для g_k^X и g^{X_1}, 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	$spc(S_1, S_2)$	$\lvert B^C \rvert$	$\lvert B_1 \backslash B^C \rvert$	$\lvert B_2 \backslash B^C \rvert$
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

Пусть (V^J, I^J) — граф синтагм, J — индексное множество, на котором задаются $L(T_i) : T_i \in T^S$, (V_1^J, I_1^J) — дерево-прецедент для T^S,

$V_1^J = J$, $I_1^J = \{(j,k) : \exists E \in V^J, (j,k) \in E\}$,

$K^E = (G^E, M^E, I^E)$ есть искомый формальный контекст эталона.

Если $\exists E \in V^J : (j,k) \in E$, а дерево (V_1^J, I_1^J) расширено с учётом *леммы 3* и *теоремы 8*, то для основ b_j и b_k и флексий f_j и f_k элементы множеств G^E, M^E и отношения I^E 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 M^E включаются признаки $m_1 = p_{bs} \odot b_k$, $m_2 = p_{bf} \odot f_k$, $m_3 = p_{fl} \odot f_j$ и $m_4 = f_j \odot «:» \odot f_k$, основа b_j включается в множество G^E, $I^E = I^E \cup \{(b_j, m_1), (b_j, m_2), (b_j, m_3), (b_j, m_4)\}$.

Случай 2

Между словами с индексами j и k стоит предлог p_y.

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

Коэффициент сжатия информации по основам относительно модели СЯУ в виде формального контекста (16) равен:

$$k^S = \frac{\sum_{i=1}^{n^{BS}} k_i^S}{n^{BS}}, \tag{23}$$

где в соответствии с принятым разбиением множества M^S

$$k_i^S = \frac{\sum_{j=1}^{n_i^{BS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AS}}{n_i^{BS}},$$

$$n^{BS} = \left|M_1\right|,$$

$$n^{MF} = \left|M_2\right|,$$

$$n_i^{BS} = \left| \left\{ g \in G^S : I^S(g,m) = \text{true},\ m \in M_1,\ m = p_{bs} \odot b_i \right\} \right|,$$

$$n_{ijk}^{AS} = \left| \left\{ m_k \in M_3 : I^S(g,m_k) = \text{true}, \right.\right.$$
$$\left.\left. \exists\ m_{bf} \in M_2 : m_{bf} = p_{bf} \odot f_k,\ m_k = b_i \odot «:» \odot f_k \right\} \right|.$$

Коэффициент сжатия по флексиям аналогичен коэффициенту по основам:

$$k^F = \frac{\sum_{i=1}^{n^{FS}} k_i^F}{n^{FS}}, \tag{24}$$

где

$$k_i^F = \frac{\sum_{j=1}^{n_i^{FS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AF}}{n_i^{FS}},$$

$$n^{FS} = \left|M_5\right|,$$

$$n^{MF} = \left|M_2\right|,$$

$$n_i^{FS} = \left|\left\{ g \in G^S : I^S(g, m) = \text{true},\ m \in M_5,\ m = p_{fl} \odot f_i \right\}\right|,$$

$$n_{ijk}^{AF} = \left|\left\{ m \in M_4 : I^S(g_j, m) = \text{true},\ \exists\, m_{bf} \in M_2 : m_{bf} = p_{bf} \odot f_k,\ m = f_i \odot «:» \odot f_k \right\}\right|.$$

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

Утверждение 7

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

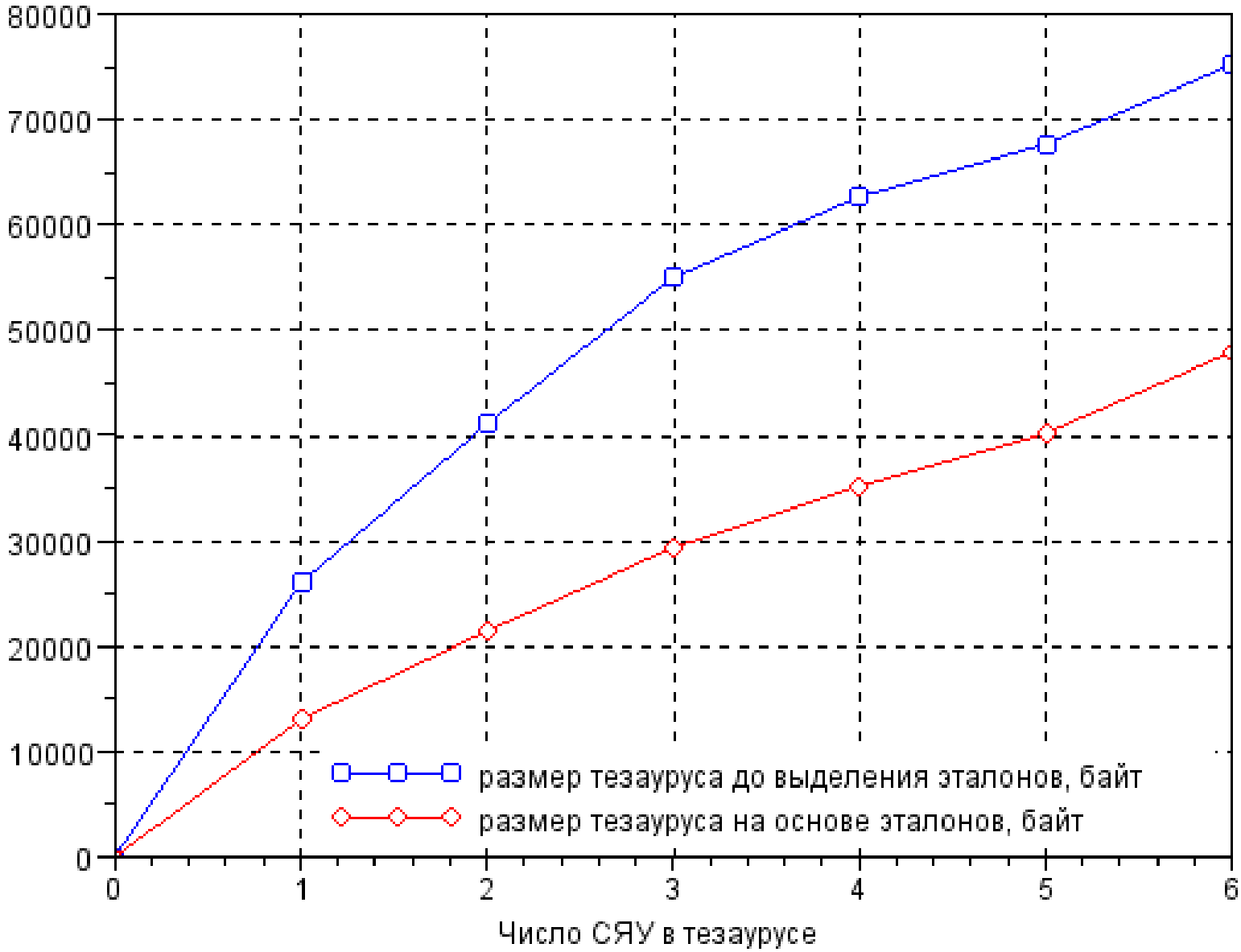
Последовательности из трех и более соподчиненных слов, встречающиеся более чем в 49% исходных СЭ-фраз, выделяются на этапе синтаксического разбора. Для каждой из них строится отдельный формальный контекст, идентичный по структуре формальному контексту эталона.

Ситуации языкового употребления и смысловые эталоны

Порядковый номер СЯУ, i	1	2	3	4	5	6
Число фраз, задающих СЯУ	54	53	26	26	2	3
из них представляют эталон	14	15	5	11	2	3
Исходное число объектов СЯУ	13	15	13	12	8	11
Исходное число признаков СЯУ	160	153	135	102	46	68
Число объектов эталона	9	12	12	12	8	11
Число признаков эталона	75	78	65	71	46	68

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



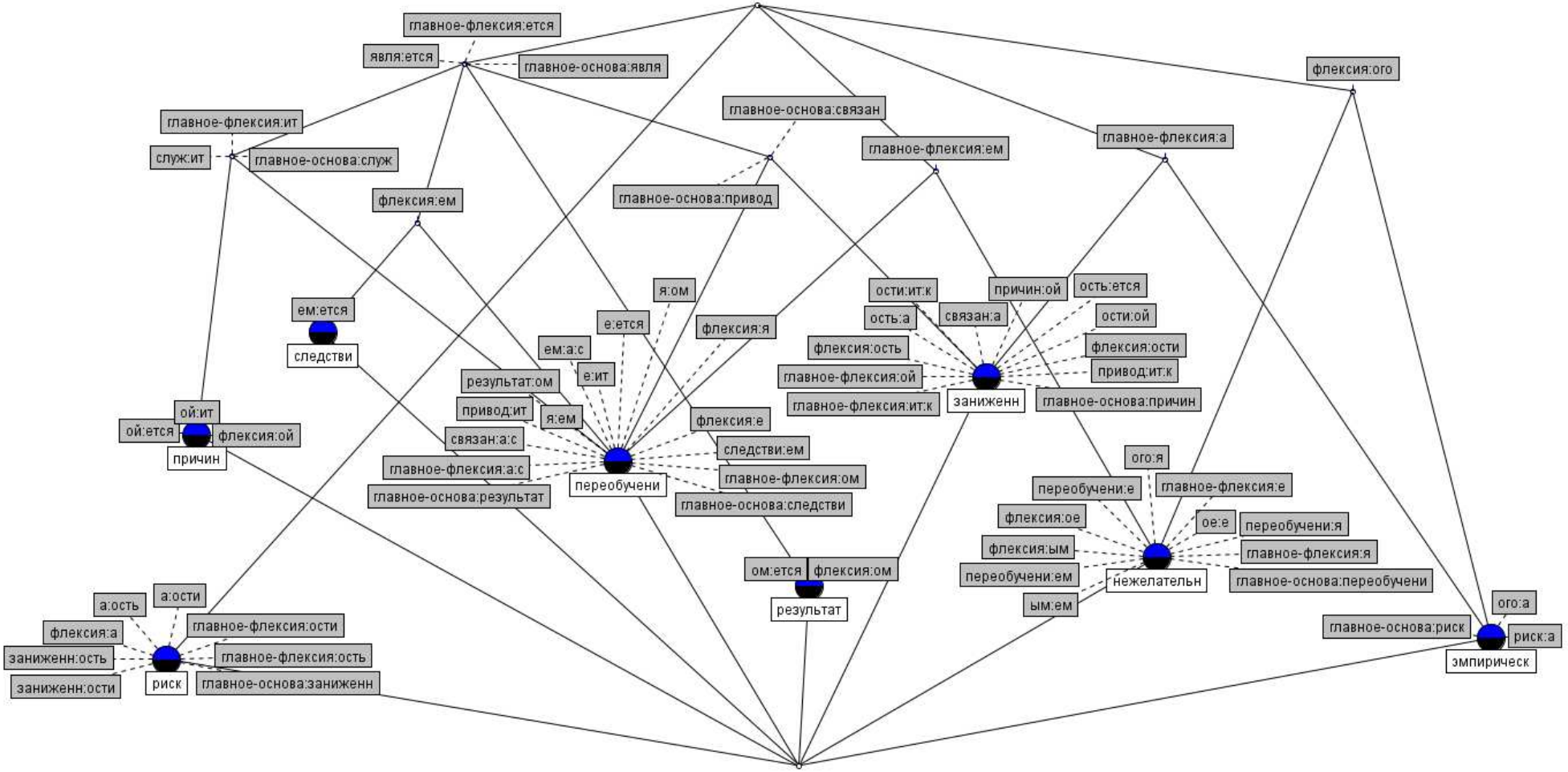
Пример: исходное множество семантически эквивалентных фраз

Синонимичные перифразы

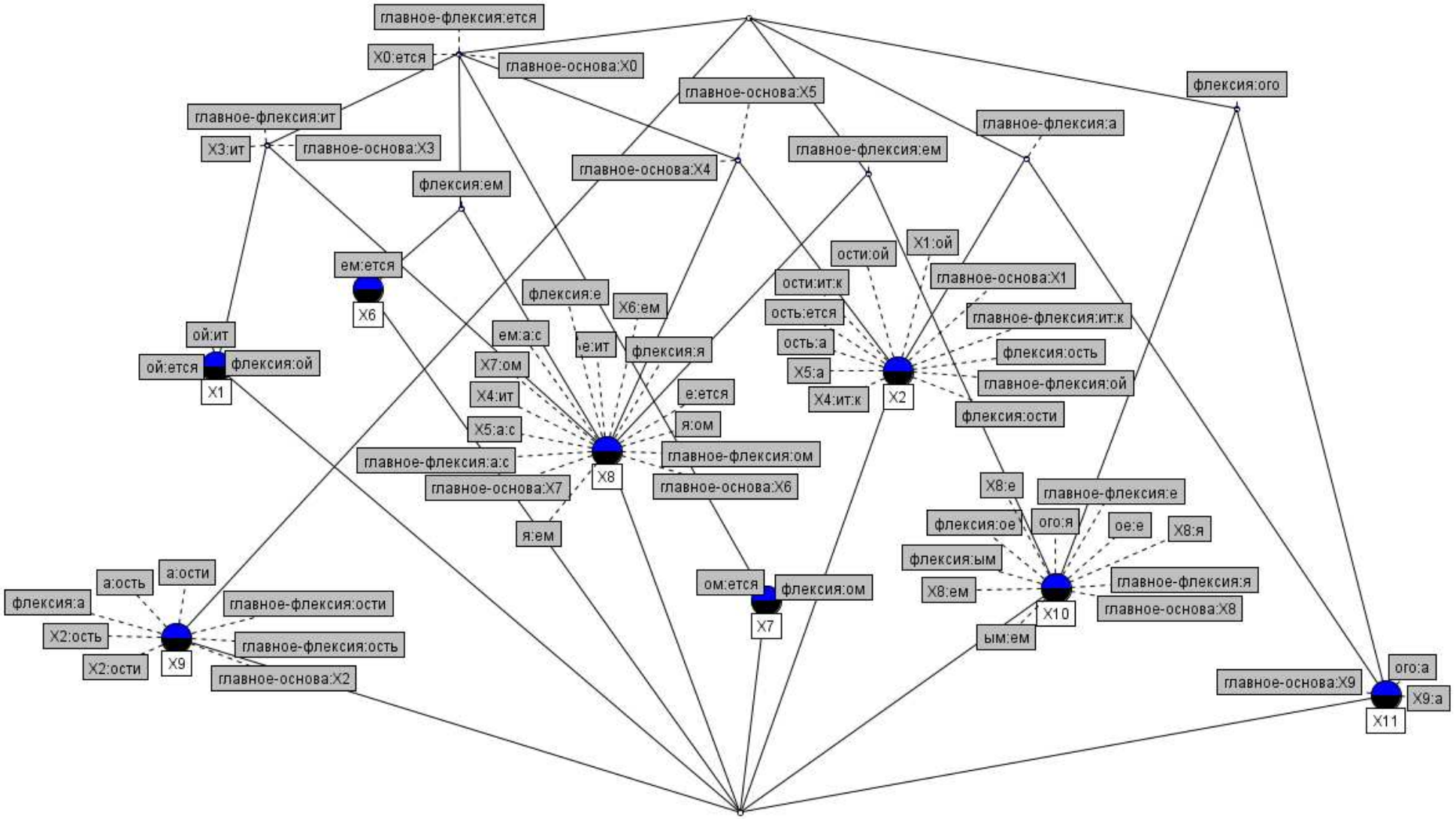
27:89 Insert Indent Modified

"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



Шаблон формального контекста эталона



Разметка СЭ-фраз в рамках шаблона СЯУ

Разметка СЭ-фраз в составе шаблона СЯУ

11:177 Insert Indent Modified

```
[[wm("X10","ое"),wm("X8","е"),wm("X4","ит"),wm("к"," "),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X10","ое"),wm("X8","е"),wm("X6","ем"),wm("которого"," "),wm("X0","ется"),wm("X2","ость"),wm("X11","ого"),wm("X9",
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ется"),wm("X6","ем"),wm("X10","ого"),wm("X8","я")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ющаяся"),wm("X6","ем"),wm("X10","ого"),wm("X8","я")],
 [wm("X11","ий"),wm("X9"," "),wm("X2","ость"),wm("которого"," "),wm("X0","ется"),wm("X6","ем"),wm("X10","ого"),wm("X8",
 [wm("X11","ий"),wm("X9"," "),wm("X2","ый"),wm("вследствие"," "),wm("X10","ого"),wm("X8","я")],
 [wm("X11","ий"),wm("X9"," "),wm("к"," "),wm("X2","ости"),wm("которого"," "),wm("ведет"," "),wm("X10","ое"),wm("X8","е")],
 [wm("X9"," "),wm("X2","ый"),wm("как"," "),wm("X6","е"),wm("X8","я")],
 [wm("X11","ий"),wm("X9"," "),wm("по"," "),wm("X1","е"),wm("обусловленной"," "),wm("X10","ым"),wm("X8","ем"),wm("може
 [wm("X11","ий"),wm("X9"," "),wm("в"," "),wm("силу"," "),wm("обстоятельств"," "),wm("X5","ных"),wm("с"," "),wm("X10","ым"
                                                                                                  wm("может","
 [wm("X11","ий"),wm("X9"," "),wm("по"," "),wm("X1","е"),wm("вызванной"," "),wm("X10","ым"),wm("X8","ем"),wm("может","
 [wm("X11","ий"),wm("X9"," "),wm("к"," "),wm("X2","ости"),wm("которого"," "),wm("X4","ит"),wm("X10","ое"),wm("X8","е")],
 [wm("X10","ое"),wm("X8","е"),wm("X3","ит"),wm("X1","ой"),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X1","ой"),wm("которой"," "),wm("X0","ется"),wm("X10","ое"),wm("X8",
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ется"),wm("X7","ом"),wm("X10","ого"),wm("X8","я")],
 [wm("X10","ое"),wm("X8","е"),wm("с"," "),wm("которым"," "),wm("X5","а"),wm("X2","ость"),wm("X11","ого"),wm("X9","а")],
 [wm("X11","ий"),wm("X9"," "),wm("с"," "),wm("X8","ем"),wm("X5","а"),wm("его"," "),wm("X2","ость")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X5","а"),wm("с"," "),wm("X8","ем")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ющаяся"),wm("X7","ом"),wm("X10","ого"),wm("X8","я")],
 [wm("X10","ое"),wm("X8","е"),wm("X7","ом"),wm("которого"," "),wm("X0","ется"),wm("X2","ость"),wm("X11","ого"),wm("X9",
 [wm("X10","ое"),wm("X8","е"),wm("X7"," "),wm("которого"," "),wm("есть"," "),wm("X2","ость"),wm("X11","ого"),wm("X9","а")
 [wm("X10","ое"),wm("X8","е"),wm("X4","ящее"),wm("к"," "),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X10","ое"),wm("X8","е"),wm("X3","ащее"),wm("X1","ой"),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("относится"," "),wm("к"," "),wm("X6","ю"),wm("X10","ого"),wm("X8","я")]
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X5","а"),wm("с"," "),wm("X10","ым"),wm("X8","ем")],
 [wm("X10","ое"),wm("X8","е"),wm("X0","ется"),wm("X1","ой"),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X1","ой"),wm("которой",""),wm("X3","ит"),wm("X10","ое"),wm("X8","е")
```

Конкретизации переменных

X_i	основа	X_i	основа	X_i	основа
X_0	явля	X_4	привод	X_8	переобучени
X_1	причин	X_5	связан	X_9	риск
X_2	заниженн	X_6	следстви	X_{10}	нежелательн
X_3	служ	X_7	результат	X_{11}	эмпирическ

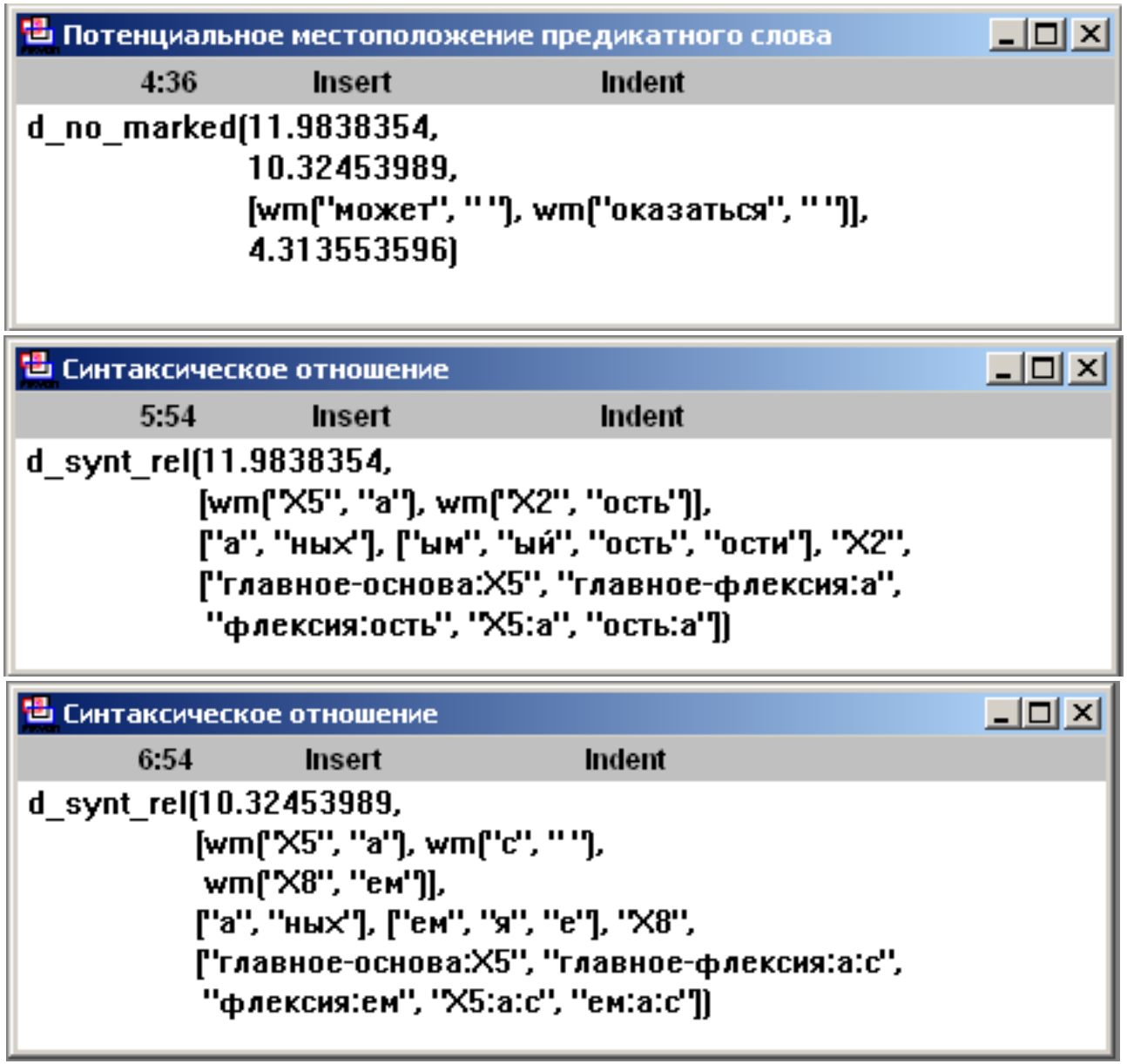
Пусть в формальном контексте эталона все обозначения основ в именах объектов и признаков заменены переменными, для каждой из которых задана конкретизация некоторой основой.

Положим аналогичные замены выполняемыми для каждой из исходных СЭ-фраз с формированием пары «основа–флексия» для каждого слова, множество последовательностей указанных пар обозначим как T^P.

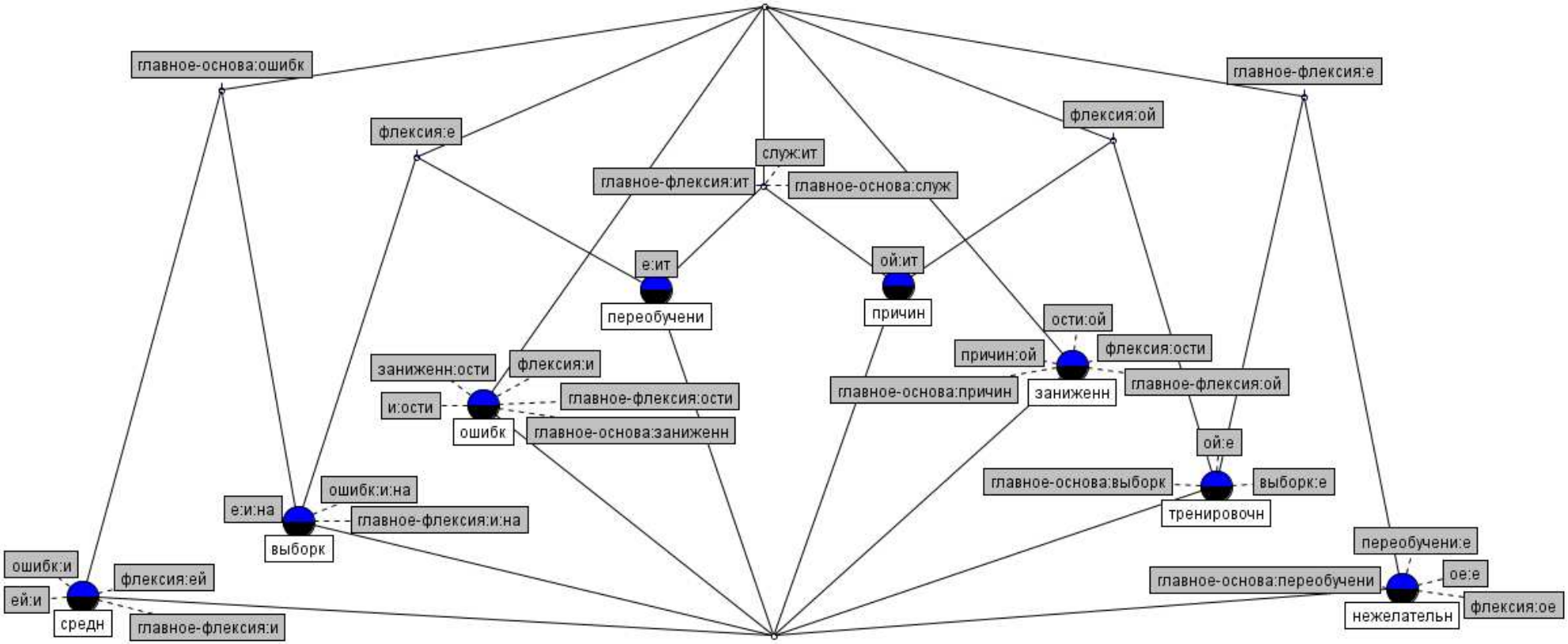
Тогда интерпретация ответа на ТЗОФ в значительном числе случаев есть «наложение» на элементы T^P, формирование списков конкретизаций и сравнение с аналогичными списками для «правильного» ответа.

Процесс происходит за линейное время, пропорциональное $\left|T^P\right|$.

Синтаксические отношения и местоположение предикатного слова



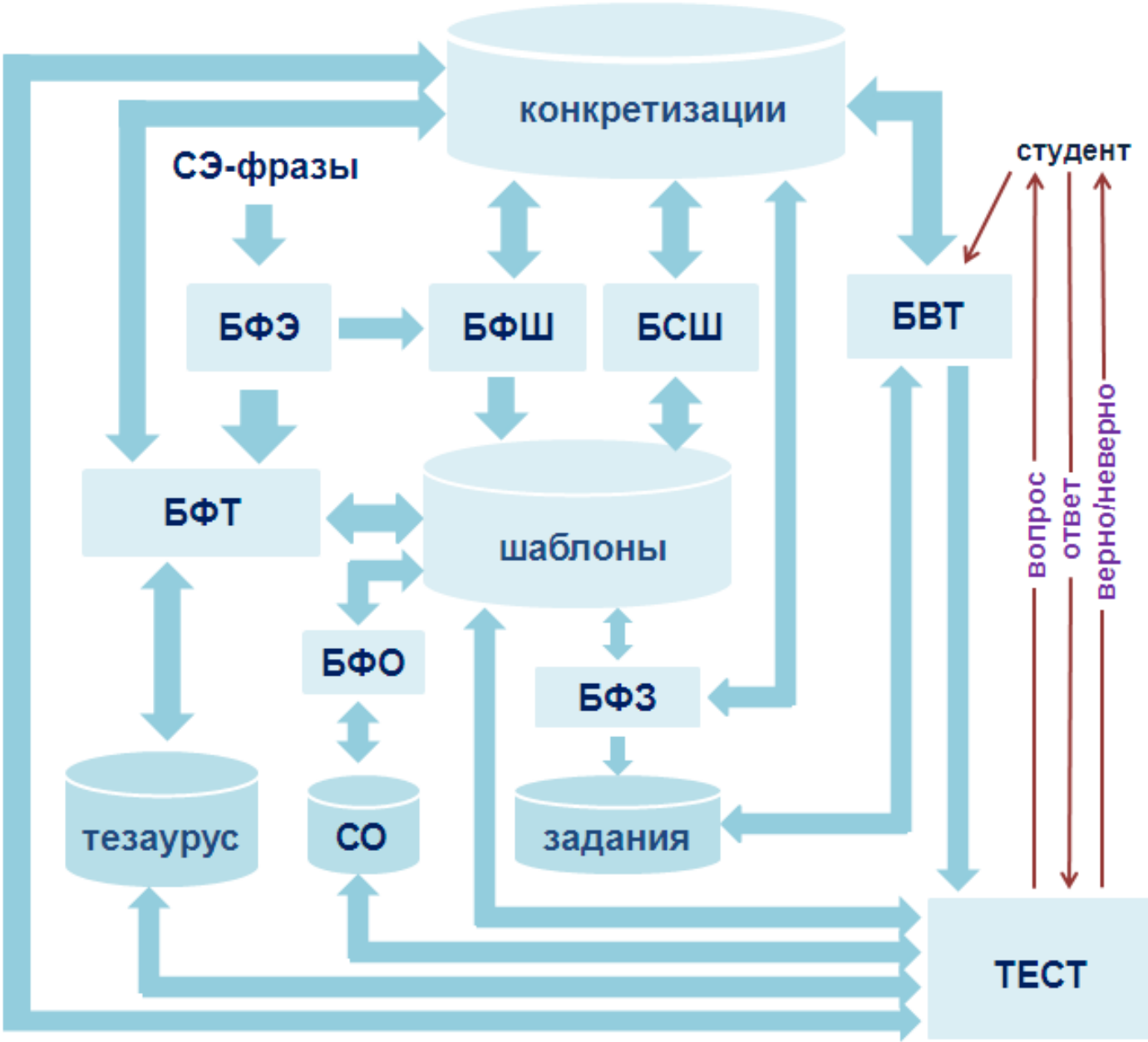
Формальный контекст смыслового эталона по результатам компиляции шаблонов двух ситуаций языкового употребления



Анализируемая фраза: «Нежелательное переобучение служит причиной заниженности средней ошибки на тренировочной выборке».

Использованы синтаксические отношения в рамках фраз: «Переусложнение модели служит причиной заниженности средней ошибки на тренировочной выборке» и «Нежелательное переобучение служит причиной заниженности эмпирического риска».

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.

Основные результаты, выносимые на защиту

1. Комплексная методика пополнения лингвистических информационных ресурсов из текстов и выделения классов смысловой эквивалентности на основе решёток формальных понятий.
2. Формальная концептуальная модель сжатия смысловой информации на основе классов смысловой эквивалентности для уровня абстрактной лексики.
3. Математическая модель и комплекс программ формирования и кластеризации семантических отношений в виде классов формальных понятий решётки и основанный на указанной модели метод выделения смыслового эталона на множестве эквивалентных по смыслу фраз предметно-ограниченного подмножества естественного языка.
4. Модель процесса интерпретации ответа на тестовое задание открытой формы распознаванием смысловых эталонов и метод компрессии текстовой базы знаний с применением указанных эталонов.
5. Метод численной оценки смысловой близости текстов предметного языка для интерпретации результатов теста открытой формы.