Автоматизация накопления знаний о синонимии и семантическая схожесть текстов предметного языка.

Михайлов Д. В., Емельянов Г. М.

Новгородский Государственный Университет имени Ярослава Мудрого

Цель работы.

Разработка и исследование методов нахождения семантического расстояния между текстами предметно-ориентированного подмножества Естественного Языка (ЕЯ).

Задачи исследования.

1) Разработка математической модели процесса формирования и кластеризации семантических отношений на основе совокупности множеств семантически эквивалентных ЕЯ-текстов заданной предметной области.
2) Определение границ проблемной области сравнения смыслов.
3) Разработка и экспериментальное исследование метода количественной оценки семантической схожести ЕЯ-текстов.
4) Выработка рекомендаций по качественному анализу предметных знаний с применением предложенных методов и моделей.

Синтагматические зависимости и ситуация языкового употребления.

Определение 1. *Ситуация Языкового Употребления (СЯУ) есть описание социального опыта человека средствами заданного Естественного Языка (ЕЯ).*

Фиксируемый при этом языковой контекст представляется тройкой:

$$S = (O, R, T),$$

где O есть множество объектов, ассоциируемых с ситуацией, R — множество отношений между $o \in O$, T — множество форм языкового описания S.

При рассмотрении $\forall\, T_i \in T$ как множества символов будет справедливым:

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

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

Пусть W_{ij} — буквенный состав слова, j — его порядковый номер в ЕЯ-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 множестве T_i^F выражаются синтагматические зависимости, которые определяют сосуществование словоформ в линейном ряду и задаются синтаксическими отношениями.

Для формирования множества R требуется на основе попарного сравнения W_{ij} различных T_i найти:

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

Синонимия и модель линейной структры предложения.

Пусть J — индексное множество для неизменных частей слов относительно T.

Определение 2. *Модель L линейной структуры предложения $T_i \in T$ есть последовательность индексов неизменных частей слов, присутствующих в T_i.*

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

Лемма 1. *Пара индексов $\{j_1, j_2\} \subset J$ соответствует словам-синонимам, если*

$$\exists \{L(T_1), L(T_2)\} \subseteq L^S : L(T_1) = J_1 \odot \{j_1\} \odot J_2 \text{ и } L(T_2) = J_1 \odot \{j_2\} \odot J_2,$$

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

Пусть P^J — множество пар, отвечающих *Лемме 1*. Преобразуем $L \in L^S$ заменой индексов, вошедших в пары из P^J, на некоторые $j \in (\mathbb{N} \backslash J)$. Обозначим преобразованное L^S как $L^{S'}$.

Теорема 1. *Индексы с максимальной встречаемостью относительно $L^{S'}$ соответствуют существительным, непосредственно подчиненным предикатному слову.*

Обозначим множество индексов, удовлетворяющих *Теореме 1*, как J^N. Пусть $L_1(T_i) \in L^{S'}$, а $L_2(T_i)$ — аналогичная модель относительно J^N. Обозначим множество моделей второго вида как L^N. Положим, что $\exists\ L_j^{S'} \subset L^{S'}$: для $\forall L_1(T_i) \in L_j^{S'}$ модели $L_2(T_i)$ одинаковы и соответствуют некоторой $L_2(T_j) \in L^N$, $T_j \in T$.

Теорема 2. *Индексы $j \notin J^N$ с максимальной частотой встречаемости в различных $L_1(T_i) \in L_j^{S'}$ соответствуют либо словам-наречиям, либо прилагательным, либо опорным существительным в составе генитивных конструкций.*

Анализ Формальных Понятий и ситуации языкового употребления.

Представим языковой контекст СЯУ посредством Формального Контекста (ФК):

$$K = (G, M, I),$$

где множество объектов G составляют основы слов, синтаксически подчиненных другим словам. Множество признаков M включает подмножества, обозначаемые далее посредством M с соответствующим нижним индексом и содержащие:

- указания на основу синтаксически главного слова (индекс 1);
- указания на флексию главного слова (индекс 2);
- связи «основа–флексия» для синтаксически 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).

Определение 3. *Пара множеств (A, B), называемых объемом и содержанием, образуют Формальное Понятие (ФП), если имеют место отображения:*

$$A' = \{m \in M \,|\, \forall\, g \in A\colon gIm\}\,, B' = \{g \in G \,|\, \forall\, m \in B\colon gIm\},$$

где $A' = B,\ B' = A,\ I \subseteq G \times M$ ставит в соответствие объектам их признаки.

Определение 4. *Множество $\Re(G, M, I)$ всех ФП формального контекста вместе с отношением порядка называется решеткой Формальных Понятий.*

Определение 5. *ФП вида (g'', g') называется объектным ФП, аналогично ФП вида (m', m'') считается признаковым ФП, где $g \in G, m \in M$.*

Расщепленные Предикатные Значения.

Теорема 3. *Пусть* $\{m_1, m_2, m_3\} \subset M_1$. *Если считать признаки* m_1, m_2 *и* m_3 *взаимно различными, то* m_1 *соответствует указанию на основу главного,* m_2 — *зависимого слова Расщепленного Предикатного Значения (РПЗ),* m_3 — *однословного смыслового эквивалента этого РПЗ при выполнении трех условий:*

1) $\exists\, g_1 \in G$: $I(g_1, m_1) = \text{true},\ I(g_1, m_3) = \text{false},\ m_2 = p_{bs} \odot g_1$;

2) $\exists\, \{g_2, g_3\} \subset G$, *при этом объекты* g_1, g_2 *и* g_3 *взаимно различны, а*

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

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

Замечание 1. *После удаления информации РПЗ формальный контекст СЯУ отражает классы отношений, которые определяются исключительно ролями объектов-участников ситуации по отношению к ней самой.*

Замечание 2. *Слова, являющиеся синонимами по Лемме 1, могут обозначать понятия с различной степенью абстракции. Указанная степень тем более, чем больше количество СЯУ, относительно которых понятие фигурирует в некоторой фиксированной роли.*

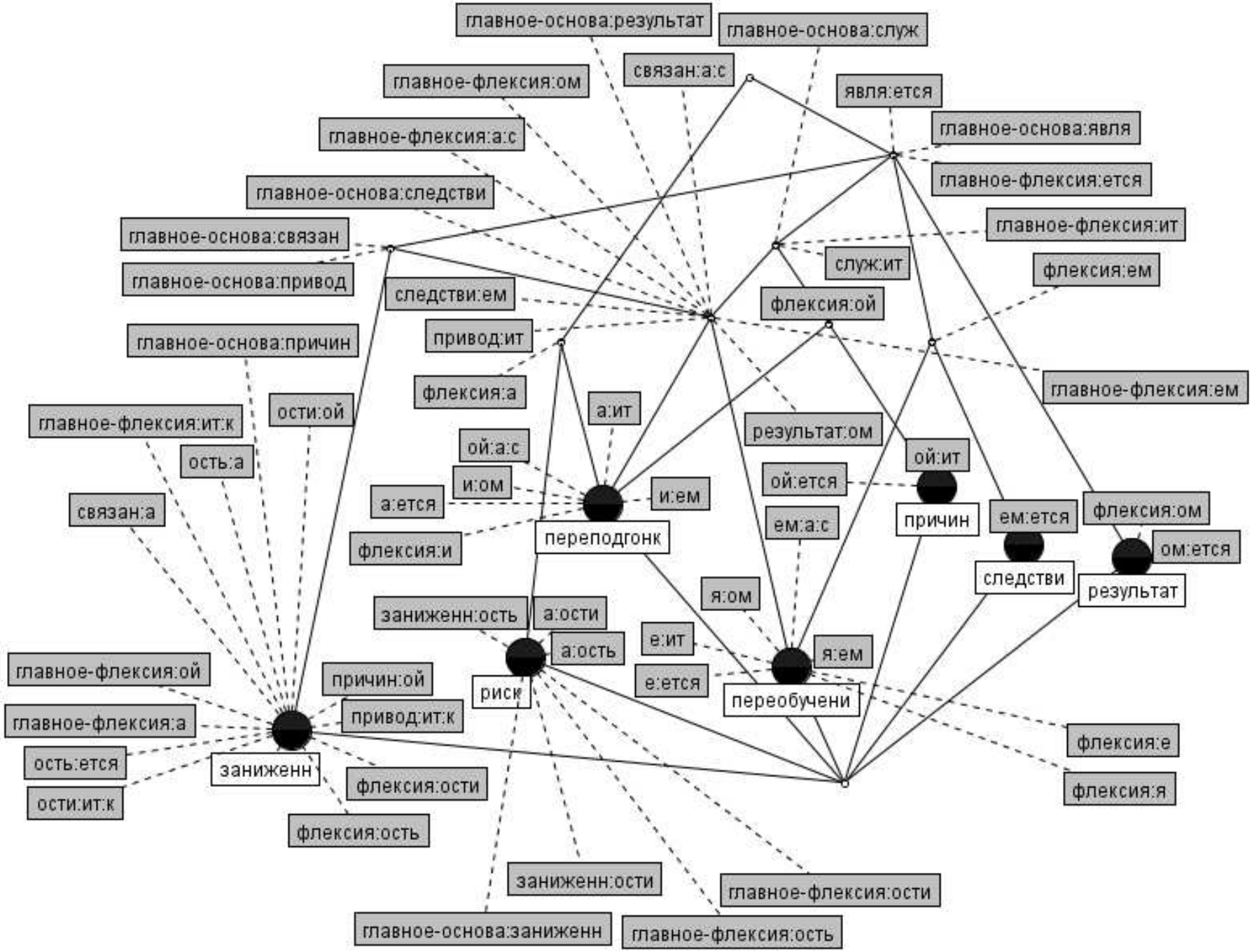

Рис. 1. Пример формального контекста СЯУ до удаления информации РПЗ.

Формальный контекст из примера на рис.1 после удаления информации расщепленных предикатных значений.

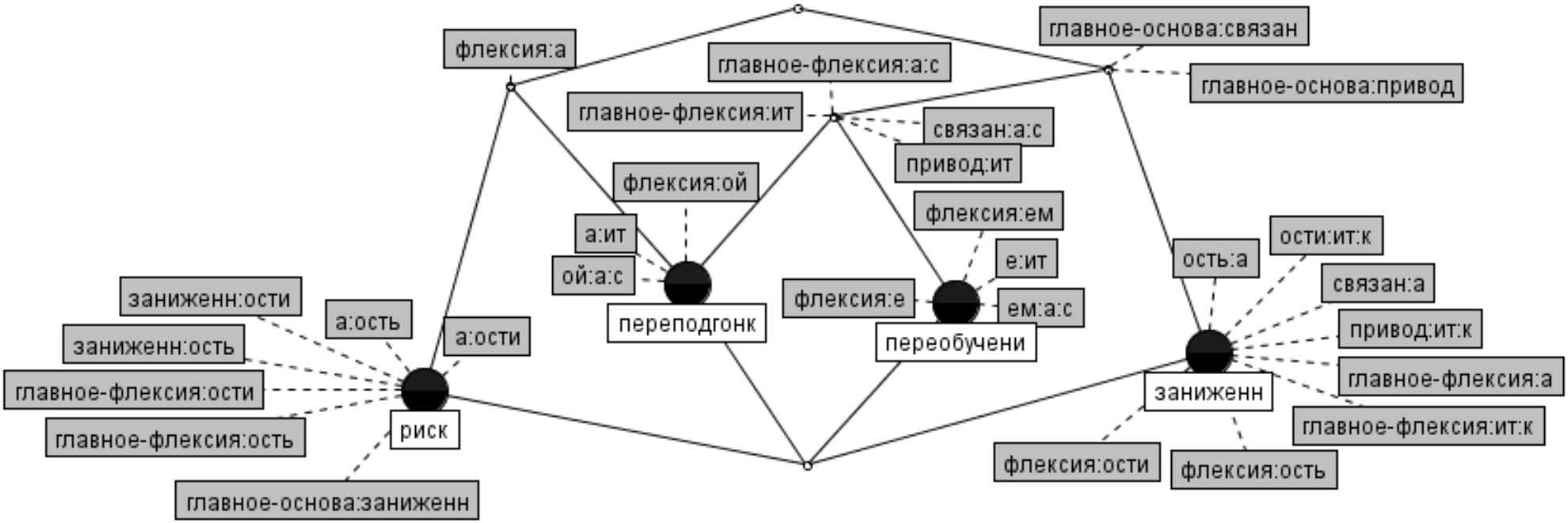

Рис. 2. Решетка ФП для редуцированного формального контекста.

Формирование тезауруса на основе совокупности СЯУ.

Рассмотрим модель тезауруса в виде формального контекста:

$$K^H = (G^H, M^H, I^H),$$

где G^H состоит из пометок отдельных СЯУ. Множество M^H содержит элементы множеств признаков ФК всех $g^H \in G^H$. Кроме того, в составе M^H выделяются:

- M_6 — множество указаний на объекты ФК отдельных $g^H \in G^H$;
- M_7 — множество связей «основа–флексия» для синтаксически зависимого слова;
- M_8 — множество сочетаний основ зависимого и главного слова.

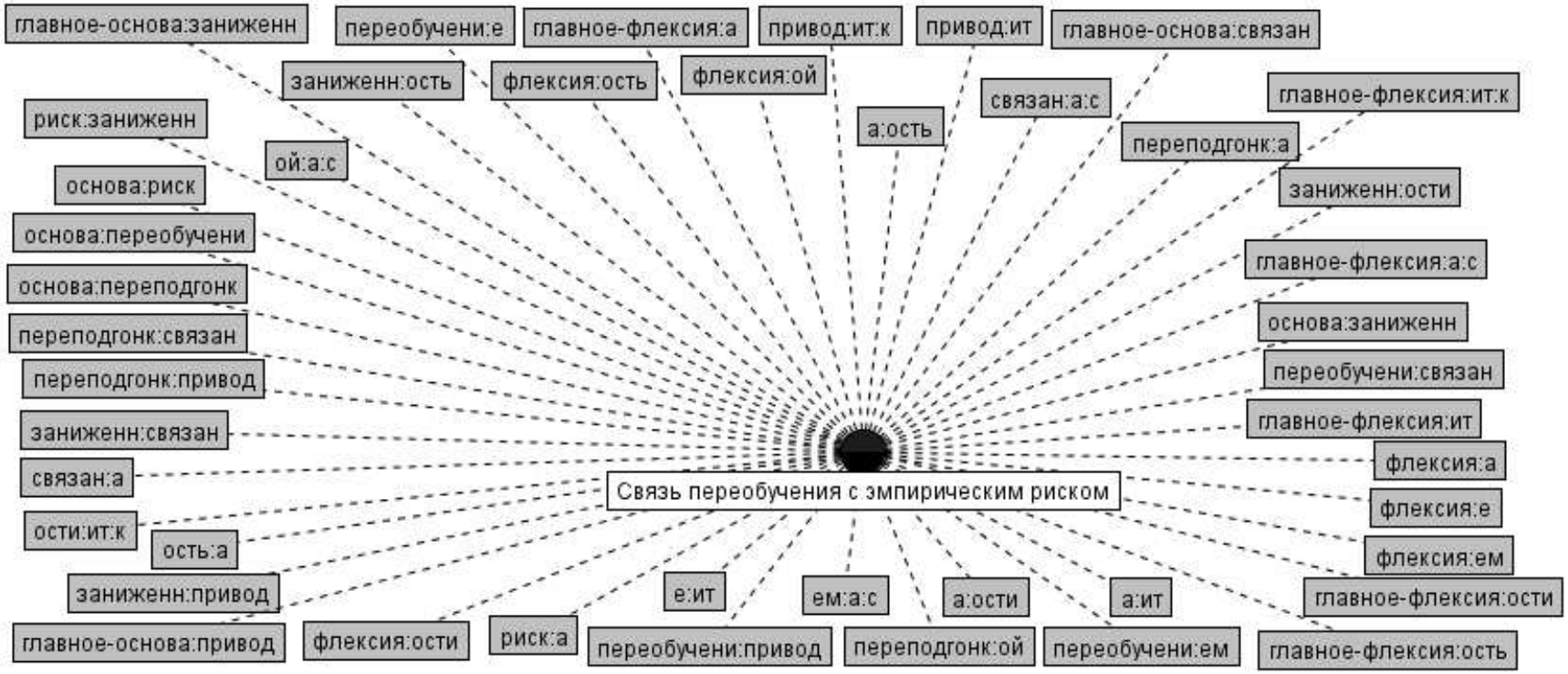

Рис. 3. Объект $g^H \in G^H$ для формального контекста на рис. 2.

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

Введем обозначения для символьных констант: p_{fl} — «флексия:», p_{bs} — «главное-основа:», p_b — «основа:», а для операции конкатенации — символ $\odot$.

Определение 6. *Будем считать, что S_1 и S_2 связаны отношением схожести, если каждому объекту $g^X \in G^X$ соответствует такой объект $g^E \in G^E$, что выполняется одно из условий:*

1) $g^X = g^E$ и любой признак $m^E \in M^E$ объекта g^E будет относиться и к объекту g^X.

2) $g^X = g^E$, при этом Условие (1) не выполняется, но существует объект $g^H \in G^H$, обладающий признаком $m_1^H \in M_6$: $m_1^H = p_b \odot g^E$ при обязательном выполнении следующих условий:

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

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

$$\left(\exists\, m_{bs}^E \in M_1^E \colon m_{bs}^E = p_{bs} \odot b^E\right) \rightarrow \left(\exists\, m_{18}^H \in M_8 \colon m_{18}^H = g^E \odot «:» \odot b^E\right), \text{при этом } I^E\left(g^E, m_{bs}^E\right) \rightarrow I^H\left(g^H, m_{18}^H\right);$$

$$\left(\exists\, m_{bs}^X \in M_1^X \colon m_{bs}^X = p_{bs} \odot b^X\right) \rightarrow \left(\exists\, m_{28}^H \in M_8 \colon m_{28}^H = g^E \odot «:» \odot b^X\right), \text{при этом } I^X\left(g^E, m_{bs}^X\right) \rightarrow I^H\left(g^H, m_{28}^H\right).$$

Кроме того, для $\forall\, m^H \in \left(M^H \backslash M^U\right)$ верно:

$$I^H\left(g^H, m^H\right) \rightarrow \left(I^E\left(g^E, m^H\right) \wedge I^X\left(g^E, m^H\right)\right). \tag{1}$$

3) $g^X \neq g^E$, но существует объект $g^H \in G^H$, обладающий признаками $m_1^H \in M_6$: $m_1^H = p_b \odot g^E$ и $m_2^H \in M_6$: $m_2^H = p_b \odot g^X$, при этом для любого признака $m^H \in \left(M^H \backslash M^U\right)$ справедливо:

$$I^H\left(g^H, m^H\right) \rightarrow \left(I^E\left(g^E, m^H\right) \wedge I^X\left(g^X, m^H\right)\right). \tag{2}$$

4) $g^X \neq g^E$, но существует $g_1^H \in G^H$, обладающий признаком $m_1^H \in M_6$: $m_1^H = p_b \odot g^E$, а для $\forall\, m^E \in \left(M_4^E \cup M_5^E\right)$ верно то, что $\left(I^H\left(g_1^H, m_1^H\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^H\left(g_1^H, m^E\right)$. При этом имеются признаки $m_2^H \in M_6$ и $m^X \in \left(M_1^X \cup M_2^X \cup M_3^X\right)$, для которых $\left(I^H\left(g_1^H, m_2^H\right) \wedge I^X\left(g^X, m^X\right)\right) \rightarrow I^H\left(g_1^H, m^X\right)$, где $m_2^H = p_b \odot g^{X_1}$, $g^{X_1} \neq g^X$, а пара $\left(g^{X_1}, g^E\right)$ отвечает Условию (3) настоящего Определения при генерации формального контекста для СЯУ g_1^H. В то же время существует объект $g_2^H \in G^H$, относительно которого пара $\left(g^X, g^{X_1}\right)$ также будет отвечать Условию (3) настоящего Определения. Генерируемый при этом формальный контекст для СЯУ g_2^H будем обозначать далее как K^{X_1}. По аналогии с K^E и K^X, $K^{X_1} = \left(G^{X_1}, M^{X_1}, I^{X_1}\right)$.

Мера схожести ситуаций языкового употребления.

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

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{3}$$

где $n = |G^X|$, а spc_k есть мера мера схожести объектов в паре (g_k^X, g^E). В зависимости от выполнимости условий *Определения 6*, значение spc_k:
- равно 1.0, если выполнено *Условие (1)*;
- вычисляется по формуле:

$$-\log_2\left(1 - \frac{D_c}{path_C}\right) \times \frac{|B^C|}{|B_1 \backslash B^C| + |B_2 \backslash B^C| + |B^C|}, \tag{4}$$

если выполнено *Условие (2), (3), либо (4).*

В случае истинности любого из *Условий (2)–(4) Определения 6* значение $D_c = 2$.

При выполнении *Условия (2)* либо *(3)* число $path_C = 4$, а в множество B^C войдут признаки $m^H \in (M^H \backslash M^U)$, для которых справедливо либо соотношение (1) (при выполнении *Условия (2)*), либо соотношение (2) (при выполнении *Условия (3)*). При этом

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

Выполнимость *Условия (4)* обычно проверяется в несколько итераций. В ходе каждой 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. При истинном *Условии (4)*

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

где $\left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right) \subset M^{X_1}$. Множество B^C здесь есть пересечение B_1 и B_2.

Таблица 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

Теоретико-решеточное представление тезауруса.

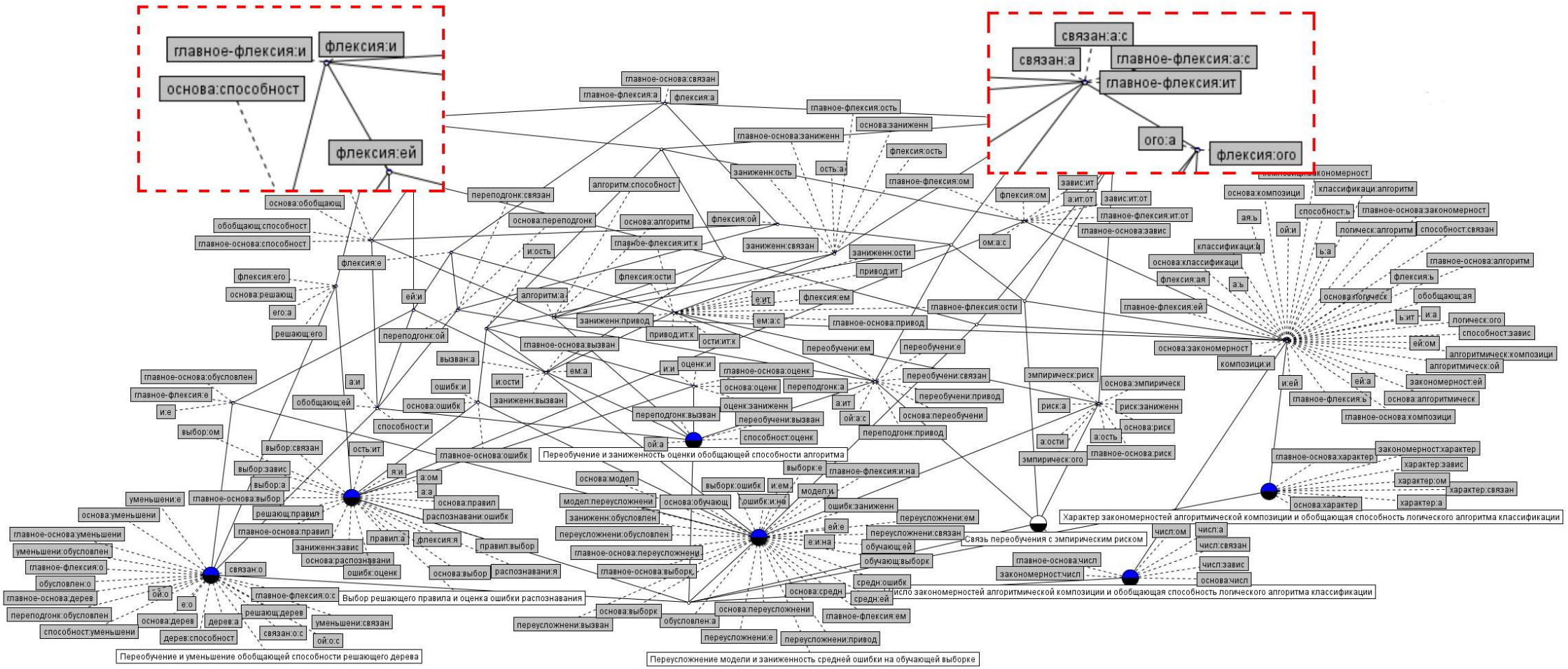

Рис. 4. Решетка ФП тезауруса и классы синтаксических отношений.

Пример : исходные данные для построения формальных контекстов сравниваемых ситуаций языкового употребления.

Таблица 2. Описание факта связи между переобучением и эмпирическим риском.

ЕЯ-описание	эталонное				анализируемые		
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

Формальный контекст ситуации языкового употребления для заведомо корректного описания заданного факта.

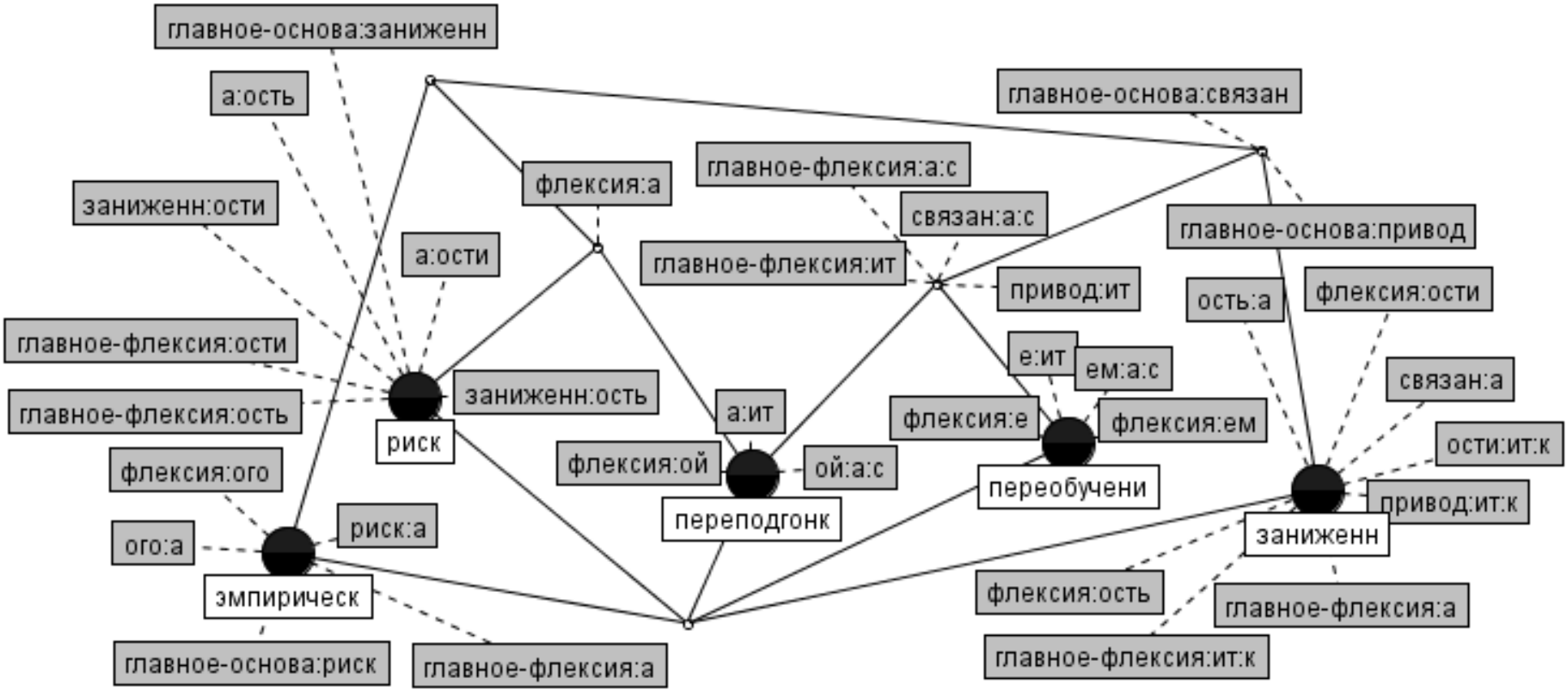

Рис. 5. Ситуация языкового употребления для «эталонного» описания факта связи между переобучением и эмпирическим риском.

Формальный контекст для Варианта 1 анализируемого описания заданного факта.

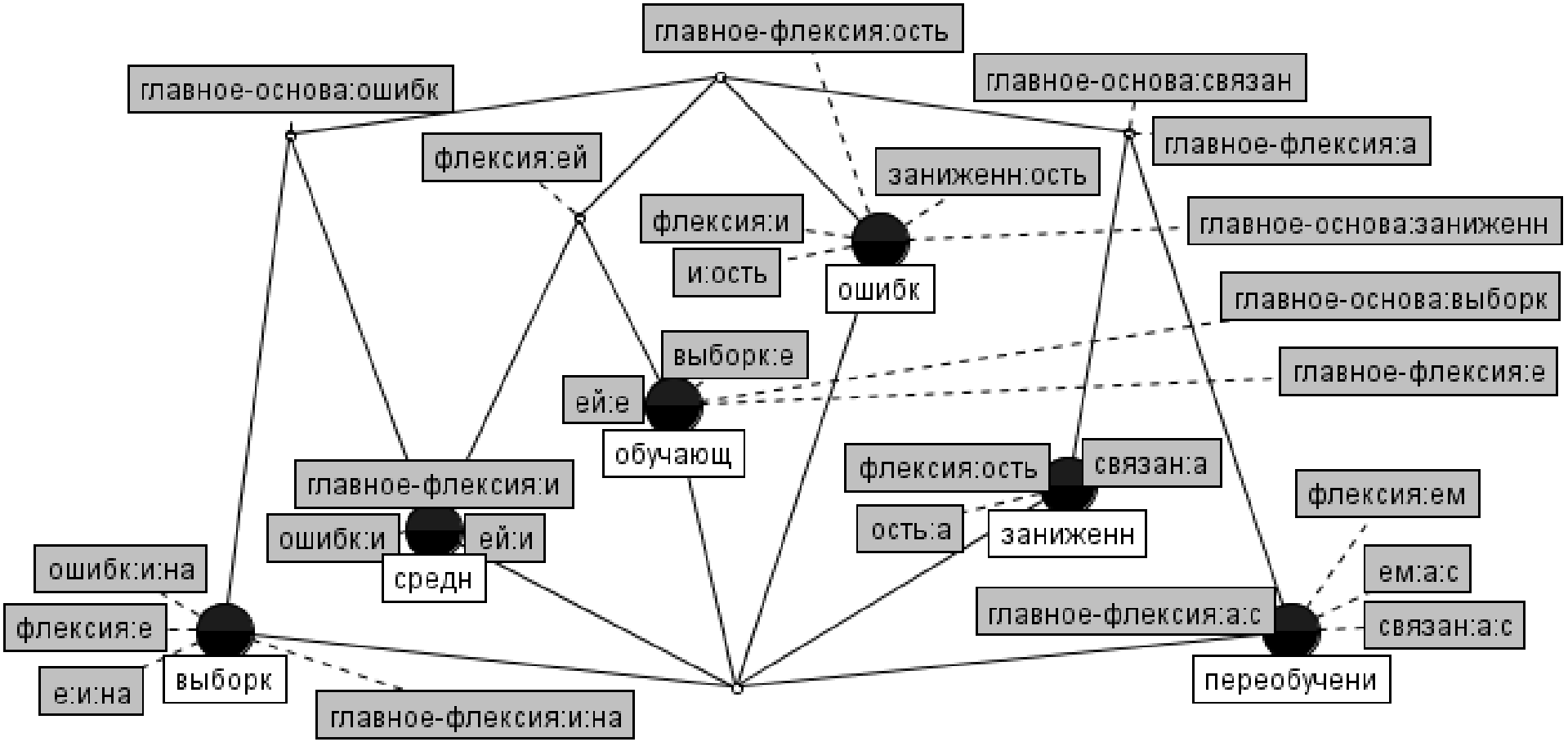

Рис. 6. Вариант 1 ЕЯ-описания связи между переобучением и эмпирическим риском.

Формальный контекст для Варианта 2 анализируемого описания заданного факта.

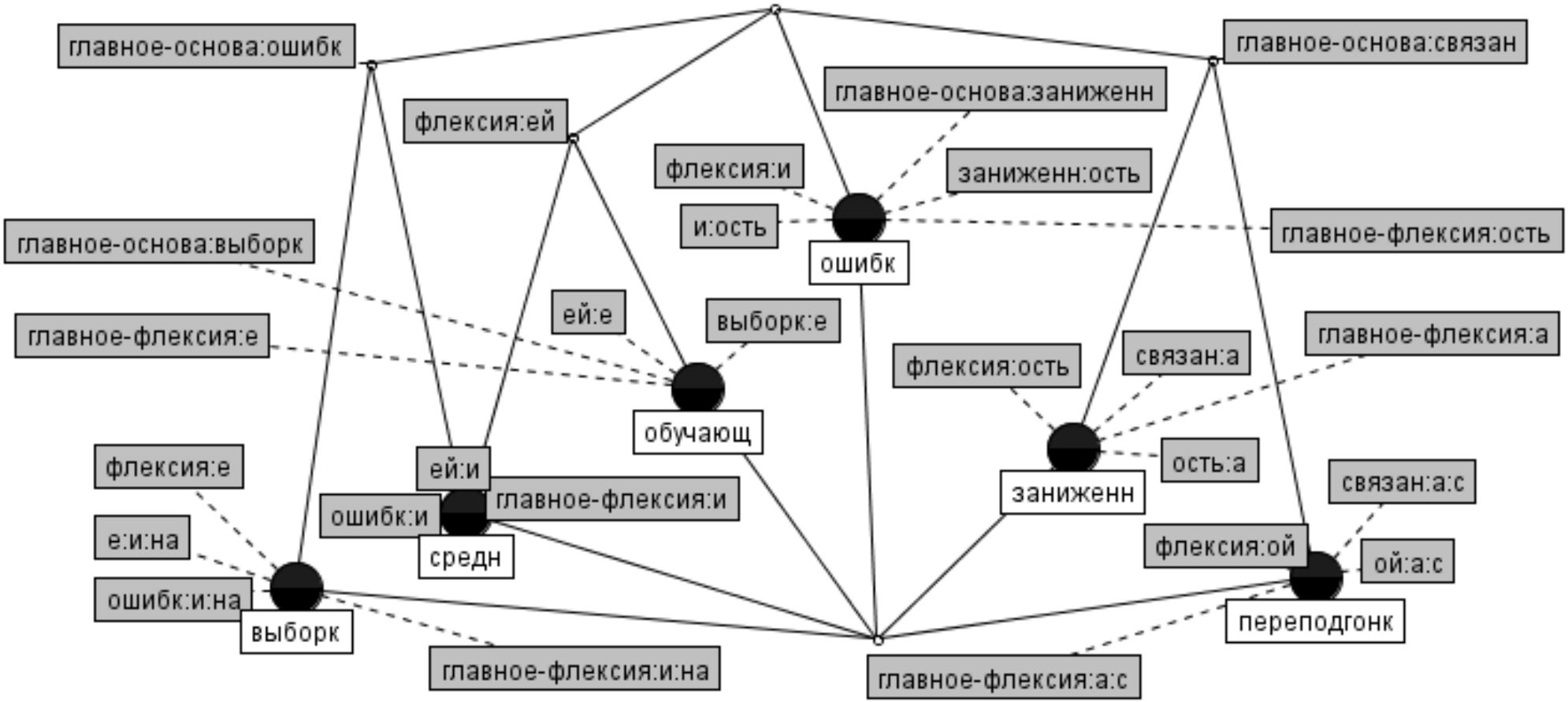

Рис. 7. Вариант 2 ЕЯ-описания связи между переобучением и эмпирическим риском.

Формальный контекст для Варианта 3 анализируемого описания заданного факта.

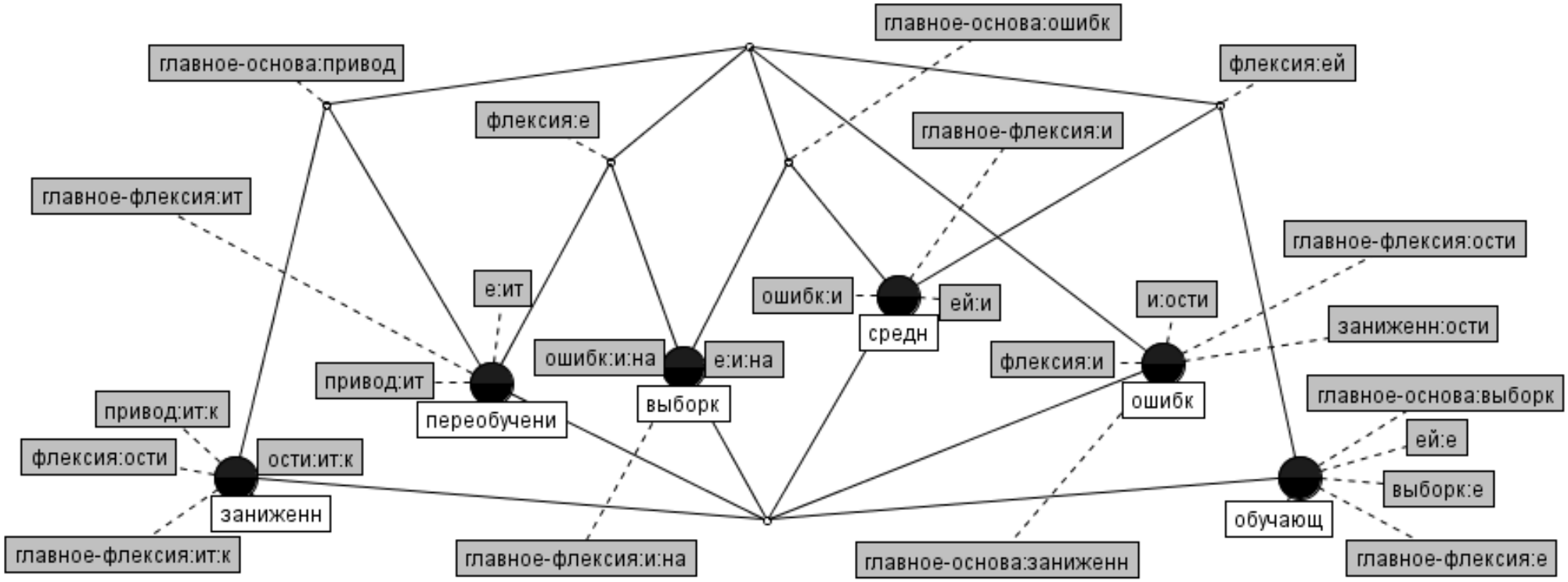

Рис. 8. Вариант 3 ЕЯ-описания связи между переобучением и эмпирическим риском.

Результат : значения близости эталону для анализируемых вариантов описания заданного факта предметной области.

Таблица 3. Сравнение вариантов ЕЯ-описания связи между переобучением и эмпирическим риском.

Вариант	$spc(S_1, S_2)$	$\left\|B^{LCS}\right\|$	$\left\|B_1 \backslash B^{LCS}\right\|$	$\left\|B_2 \backslash B^{LCS}\right\|$
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

Выводы.

- Основной *результат* настоящей работы — *метод анализа схожести ситуаций языкового употребления при их независимом порождении*. Сфера применения предложенного *метода* — задачи семантического анализа, для которых заранее неизвестно соответствие сравниваемых ЕЯ-высказываний тезаурусной информации. Унифицируемое теоретико-решеточное представление сравниваемых высказываний и тезаурусной информации позволяет максимально просто пополнять тезаурус и эффективно использовать имеющуюся в нем информацию при анализе близости текстов.

- Предложенная *модель тезауруса* может быть использована в качестве основы построения текстовых баз данных для заданной предметной области. Организация текстовой базы данных на основе решетки Формальных Понятий позволяет за счет иерархического представления информации уменьшить как размер самой базы данных, так и время поиска в ней.

- *Отдельного обсуждения* заслуживает интеграция предложенного метода с лингвистическими и статистическими методами информационного поиска, используемыми алгоритмом Exactus, http://www.exactus.ru/.