

УРОВЕНЬ ГЛУБИННОГО СИНТАКСИСА И ВЫДЕЛЕНИЕ СВЕРХФРАЗОВЫХ ЕДИНСТВ В ТЕКСТАХ ПРИ УСТАНОВЛЕНИИ ИХ СЕМАНТИЧЕСКОЙ ЭКВИВАЛЕНТНОСТИ

Цель состоит в разработке и исследовании математической модели процесса распознавания и построения формальных семантических образов сверхфразовых единств в высказываниях на Естественном Языке (ЕЯ) при установлении их семантической эквивалентности.

Задачи:

- 1) Построение концептуальной модели процесса распознавания смысловой взаимной дополняемости фраз анализируемого высказывания;
- 2) Исследование алгоритмической разрешимости и сложности модели;
- 3) Разработка методов и алгоритмов для программной реализации полученной модели;
- 4) Исследование вопросов взаимодействия процессов построения образов сверхфразовых единств в анализируемом тексте и установления его эквивалентности смысловому эталону.

Системообразующие свойства правил расширенной лексико-синтаксической Δ - грамматики

Определение 1. Лексическая Синонимическая Конструкция (ЛСК) – заменяемый лексическим правилом комплекс лексических единиц и связывающих их отношений глубинного синтаксиса. T_1 и T_2 – деревья Глубинных Синтаксических Структур (ГСС) фраз F_1 и F_2 :

$$\begin{aligned} T_1 &= C(T_1^0; \alpha_0^1 \mid C(t_1; \alpha_1^1, \alpha_2^1, \dots, \alpha_k^1 \mid T_1^1, T_1^2, \dots, T_1^k)) \\ T_2 &= C(T_2^0; \alpha_0^2 \mid C(t_2; \alpha_1^2, \alpha_2^2, \dots, \alpha_l^2 \mid T_2^1, T_2^2, \dots, T_2^l)), \text{ где} \\ t_1 &= C(t_1^0; \alpha_0^w \mid C(t_1^w(C_0); \alpha_1^w, \alpha_2^w, \dots, \alpha_n^w \mid t_1^1, t_1^2, \dots, t_1^n)) \\ t_2 &= C(t_2^0; \beta_0^w \mid C(t_2^w(C_0); \beta_1^w, \beta_2^w, \dots, \beta_m^w \mid t_2^1, t_2^2, \dots, t_2^m)) \end{aligned}$$

Синтаксическая замена : $t_1 \Rightarrow t_2$;

Лексическая замена : $t_1^w \Rightarrow t_2^w$; (*)

С-операция композиции;

t_1^w и t_2^w - деревья ЛСК, C_0 - ключевое слово ЛСК.

Определение.2. Деревья ГСС $\Phi 1$ и $\Phi 2$

удовлетворяют необходимому, но не достаточному условию ЛФ-синонимии, если их ЛСК относится к

одному и тому же ключевому слову C_0 .

Определение 3. Отвечающие необходимому

условию ЛФ-синонимии деревья $T1$ и $T2$

удовлетворяют необходимому (но не достаточному !) условию взаимной дополняемости (смысловой),

если существует последовательность

преобразований (*), приводящих $T1$ и $T2$ к виду с одинаковой ЛСК.

Функциональные требования к модели

- Анализ применимости каждого преобразования из заданного множества к каждому дереву в смысловом описании высказывания с представлением результата в виде списка :
 $((\text{правило_}i \ C_0(i)) \dots (\text{правило_}k \ C_0(k)))$;
- Построение последовательности преобразований для приведения удовлетворяющих необходимому условию ЛФ-синонимии деревьев ГСС к виду с одинаковой ЛСК;
- Определение наличия взаимной дополняемости* деревьев ГСС;
- Суммирование взаимно дополняющих друг друга деревьев с последующей их заменой в смысловом описании анализируемого высказывания на ГСС их суммы.

* **Определение.** Приведенные к виду с одинаковой ЛСК деревья глубинного синтаксиса $T1$ и $T2$ взаимно дополняют друг друга, если они изоморфны таким образом, что для всякого узла α дерева $T1$ его образ $f(\alpha)$ в дереве $T2$:

- либо содержит информацию об одной и той же ненулевой характеризованной обобщенной лексеме;
- либо представляет фиктивную лексему с теми же семантическими словоизменительными характеристиками, что и ненулевая характеризованная обобщенная лексема в узле α ;

Формальная концептуальная модель процесса распознавания смысловой взаимной дополняемости фраз анализируемого высказывания

Y_S - язык смыслов заданного ЕЯ Y :

$Y_S = \langle L_S, \Gamma_S, \Pi, Q, U \rangle$, где

L_S – лексика языка Y_S ,

Γ_S - синтаксис языка Y_S ,

Π – процедура установления соответствий между фразами языков Y и Y_S ,

Q – процедура установления эквивалентности в Y_S ,

U – процедура преобразования текста на основе учета семантических повторов.

$U = \langle Q_U, S_U \rangle$, где

Q_U – процедура приведения связанных по смыслу фраз в Y_S к целевому* представлению,

S_U – процедура построения суммарного смысла в языке Y_S .

* Под целевым представлением фраз в языке Y_S понимается представление, допускающее суммирование

Представление системы правил Δ -грамматики ограниченной сетью Петри

Множество $\{T^\pi\}$ входов и выходов правил $\pi \in (\Pi^R \setminus \Pi_U^R)$ произвольных элементарных преобразований в Γ_L^R составляет множество информационных элементов. $\{T^\pi\} = \{T_1^\pi\} \cup \{T_2^\pi\}$, где $\{T_1^\pi\}$ -множество “входов”, $\{T_2^\pi\}$ -множество “выходов” правил.

Сеть N_i , моделирующая работу i -й системы правил, где $i \in 1, \dots, n_sys$:

$$N_i = \{P_i, T_i, F, H, M_{0i}\}, \text{ где}$$

множество позиций P_i есть подмножество $\{T^\pi\}$ элементов, составляющих входы и выходы способных образовывать систему правил;

Множество T_i переходов сети составляет множество переходов между состояниями системы:

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

Компонент R в описании правила $\pi \in (\Pi^R \setminus \Pi_U^R)$ отвечает за его применимость и представляется логической функцией:

$$r_j = x^1 \wedge x^2 \wedge \dots \wedge x^n$$

Правило π может быть применено к дереву T_1^π , если выполняется одно из условий $r_j \in R : \bigvee_{j=1}^m r_j = true$.

Задача приведения деревьев глубинного синтаксиса к виду с одинаковой ЛСК

Применение правила $\pi \in (\Pi^R \setminus \Pi_U^R)$ сводится к выполнению перехода :

$$\pi(r_{12}) : T_1^\pi \xrightarrow{\pi(r_{12})} T_2^\pi, \text{ где } r_j = x^1 \wedge x^2 \wedge \dots \wedge x^n.$$

Последовательность применяемых правил моделируется последовательностью $\tau = (t_i^1, t_i^2, \dots, t_i^k)$ срабатываний переходов :

$$T_1^\pi \xrightarrow{\pi_1(r_{12})} T_2^\pi \xrightarrow{\pi_2(r_{23})} T_3^\pi \rightarrow \dots \rightarrow T_k^\pi \xrightarrow{\pi_k(r_{kk+1})} T_{k+1}^\pi,$$

$$\text{где } t_i^1 \Leftrightarrow \pi_1(r_{12}), t_i^2 \Leftrightarrow \pi_2(r_{23}), \dots, t_i^k \Leftrightarrow \pi_k(r_{kk+1})$$

приводящей к последовательной смене разметок :

$$M_{0i} \xrightarrow{t_i^1} M_i^1 \xrightarrow{t_i^2} M_i^2 \dots \xrightarrow{t_i^k} M_i^k$$

$$\text{где } M_{0i} \Leftrightarrow T_1^\pi, M_i^1 \Leftrightarrow T_2^\pi, \dots, M_i^k \Leftrightarrow T_{k+1}^\pi.$$

Задача приведения деревьев T_1^π и T_{k+1}^π к виду с одинаковой ЛСК включает три задачи :

- Определение достижимости M_i^k из M_{0i} есть определение наличия слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_i^k$, где T_i^* - множество всех слов в алфавите T_i ;
- Задача обратимости слова τ : если $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_i^k$, то существует ли слово $\tau' = (t_i^{k'}, t_i^{(k-1)'}, \dots, t_i^{2'}, t_i^{1'})$:
$$M_{0i} \xleftarrow{t_i^{1'}} M_i^1 \xleftarrow{t_i^{2'}} M_i^2 \dots M_i^{k-1} \xleftarrow{t_i^{k'}} M_i^k$$

где $M_{0i} \Leftrightarrow T_1^\pi, M_i^1 \Leftrightarrow T_2^\pi, \dots, M_i^k \Leftrightarrow T_{k+1}^\pi$;
- Задача определения оптимального слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_i^k$. Если $\exists \tau_1, \tau_2, \dots, \tau_l : M_{0i} \xrightarrow{\tau_1} M_i^k$, $M_{0i} \xrightarrow{\tau_2} M_i^k$ и $M_{0i} \xrightarrow{\tau_3} M_i^k$, то берется обратимое слово минимальной длины.

Свойства языка сети, моделирующей систему правил Δ -грамматики

Лемма 1. Все правила $\pi \in (\Pi^R \setminus \Pi_U^R)$ в расширенной лексико-синтаксической грамматике Γ_L^R различны.

Лемма 2. Проблема достижимости заданной разметки M_i^k из начальной M_{0i} в сети N_i разрешима.

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 $L(N_i)$ разрешима.

Теорема 4. Проблема определения оптимального слова $\tau \in T_i^* \mid M_{0i} \xrightarrow{\tau} M_i^k$ (T_i^* - множество всех слов в алфавите T_i) в языке $L(N_i)$ сети N_i является разрешимой.

Исчисление сценариев на информационном пространстве системы правил Δ -грамматики

$S_i^j \in S_i : S_i^j = \{T_k^\pi, T_l^\pi\}$, где $T_k^\pi \in \{T^\pi\}$, $T_l^\pi \in \{T^\pi\}$.

$$S_i^j = \{ref_i^j(k_i^j), P_i^j\}$$

$ref_i^j(k_i^j) = \{S_i^{j1}, \dots, S_i^{jk}\}$ - множество сценариев,

связанных с S_i^j через разрешенные в его рамках

переходы $t_i^j \in T_i$, $P_i^j \in \{p_j^1, p_j^2\}$ - множество позиций

сети N_i , активных в рамках S_i^j .

Целевое состояние системы $S_i^j = \{T_k^\pi, T_l^\pi\}$:

$$T_k^\pi = T_l^\pi \Leftrightarrow p_i^j : M_i(p_i^j) = 2$$

Теорема 5. Для каждого задаваемого над сетью N_i

сценария $S_i^j \in S_i$ можно указать максимум два

перехода $t_i^j \in T_i$ и $t_i^k \in T_i : t_i^j \neq t_i^k$ и $\exists S_i^{j1} \in S_i, S_i^{j2} \in S_i$,

$$S_i^{j1} \neq S_i^{j2}, : S_i^k \xrightarrow{t_i^j} S_i^{j1}, S_i^{j1} \xrightarrow{t_i^k} S_i^j, S_i^j \xrightarrow{t_i^k} S_i^{j2}, S_i^{j2} \xrightarrow{t_i^j} S_i^j.$$

Для формирования пути к найденному решению в

описание S_i^j вводится ссылка $ref_back_i^j$ на

сценарий, с которым S_i^j связан посредством

некоторого перехода $t_i^j \in T_i$:

$$S_i^j = \{ref_back_i^j, P_i^j\}$$

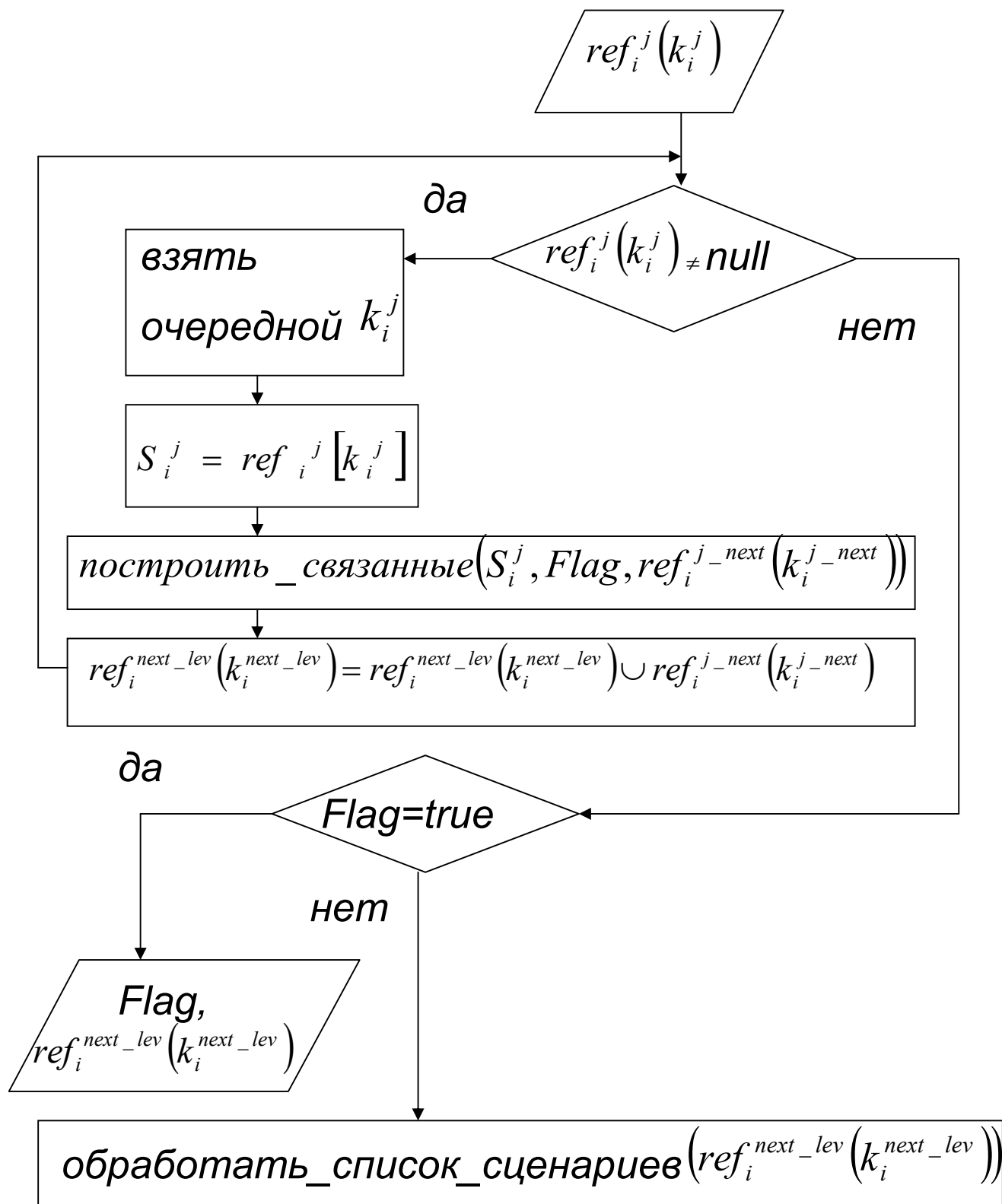
Генерация $S_i^j \in S_i$ происходит путем обработки

матрицы F с использованием массива ссылок на описания входов/выходов правил

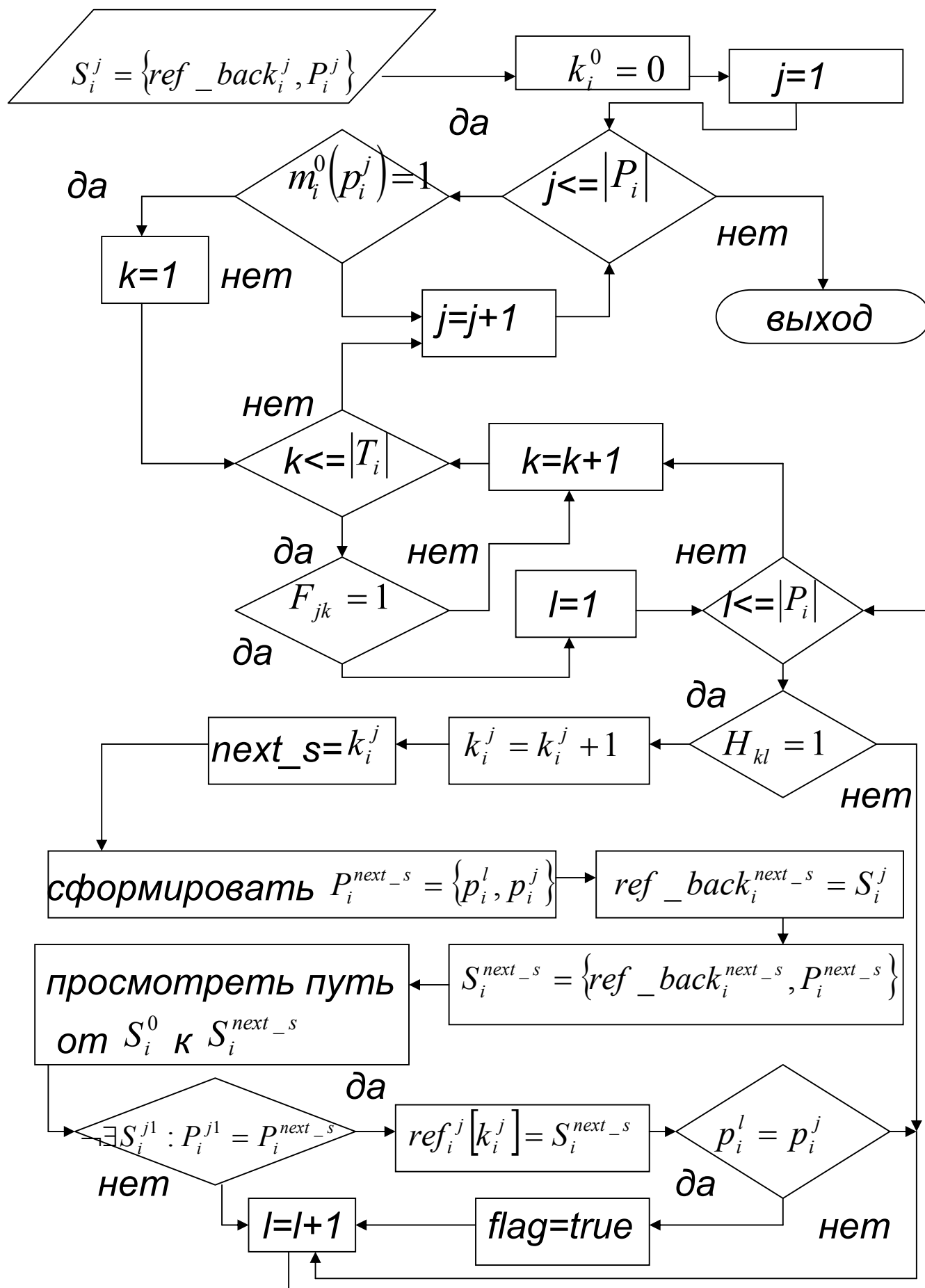
$$\Sigma_{dbfi} = P_i = \{p_i^1, p_i^2, \dots, p_i^{|P_i|}\}$$
 и описания условий

применимости $\Sigma_{Ri} = \{t_i^1, t_i^2, \dots, t_i^{|T_i|}\}$.

Построение системы целевых выводов на информационном пространстве системы правил Δ - грамматики



Построение множества сценариев с применением операции “ссылка назад”



Функциональное описание информационного наполнения дерева глубинного синтаксиса

Определение. W/V -дерево есть помеченное дерево с пометками в узлах из множества W и с пометками на ветвях из множества V .

Информационное наполнение узла $w_\chi \in W_\chi^{W/V}$:

$w_\chi = (lex_in_\chi, gram_in_\chi, arrow_label, composition_label)$, где $W_\chi^{W/V}$ - множество узлов дерева ГСС $T_\chi^{W/V}$ фразы χ ;
 lex_in_χ - лексическая часть w_χ ;

$gram_in_\chi$ - грамматическая часть w_χ ;

$arrow_label$ - пометка входящей в узел ветви;

$composition_label$ - композиционная метка узла.

$lex_in_\chi = (C_0, fun_n, \dots, fun_1)$, где

C_0 - ключевое слово ЛСК,

$fun_n, \dots, fun_1$ - символы Лексических Функций (ЛФ).

$gram_in_\chi = (part_of_speech, list_semant_categ)$, где

$part_of_speech$ - символьный атом, обозначающий часть речи;

S – существительное

Conj – союз

V – глагол

Num – числительное

A – прилагательное

P – причастие

Adv – наречие

Prep – предлог

$list_semant_categ$ - список семантически обусловленных словоизменительных категорий.

Аналогично представляется информация узла

$w_\pi \in W_\pi^{W/V}$ входного/выходного дерева $T_\pi^{W/V}$ правила π

Функционально-логическая модель входа/выхода правила Δ -грамматики

Порождаемые входом/выходом правила процессы моделируются сетью действий :

$$N_{\pi} = \{P_{\pi}, T_{\pi}, F_{\pi}, H_{\pi}, C, M_0^{\pi}\}, \text{ в которой}$$

P_{π} - множество состояний входа/выхода.

Каждому $t_{\pi}^i \in T_{\pi}$ соответствуют вычисления функций lex_in_{π} , $gram_in_{\pi}$, $arrow_label_{\pi}$ для очередного w_{π} .

Цветам маркера $\overset{5}{color\ i \in C}$ соответствуют способы использования информационного элемента;

Теорема 11. Пусть $N'_{\pi} = \{P'_{\pi}, T'_{\pi}, F'_{\pi}, H'_{\pi}, C, M_0^{\pi'}\}$ – сеть действий, с каждым из переходов $t_{\pi}^j \in T'_{\pi}$ связано вычисление логической функции. Тогда, если считать матрицы инцидентности F'_{π} и H'_{π} логическими функциями, то для того, чтобы $\forall t_{\pi}^j \in T'_{\pi}$ был достижим из начальной разметки $M_0^{\pi'}$, достаточно выполнения условия :

$$\bigwedge_{j=1}^{|T'_{\pi}|+1} \bigvee_{\substack{i=1 \\ i \neq j}}^{|P'_{\pi}|} \bigwedge_{k=1}^{|T'_{\pi}|} (f'_{\pi}[k, j] \wedge t_{\pi}^j \rightarrow h_{\pi}^{*'}[i, k] \wedge t_{\pi}^i) = true, \text{ где}$$

$\wedge, \vee, \rightarrow$ - логические функции конъюнкции, дизъюнкции и импликации, $f'_{\pi}[k, j] \in F'_{\pi}$, $h_{\pi}^{*'}[i, k] \in H_{\pi}^{*'}$ - элементы матриц инцидентности, причем расширенная матрица $H_{\pi}^{*'}$ получена из матрицы H'_{π} путем дополнения строкой $h'_{\pi}[|T'_{\pi}|+1, 1], \dots, h'_{\pi}[|T'_{\pi}|+1, |P'_{\pi}|]$ следующим образом :
 $h'_{\pi}[|T'_{\pi}|+1, j] = 1$, если $M_0^{\pi'}(p_{\pi}^j) = 1$, $h'_{\pi}[|T'_{\pi}|+1, j] = 0$, если $M_0^{\pi'}(p_{\pi}^j) = 0$.

Алгоритмическая сложность суммирования деревьев глубинного синтаксиса

Определение. W/V деревья t_1' и t_1 считаются изоморфными с точностью до функционального соответствия, если между множествами их узлов существует взаимно-однозначное соответствие так, что в дереве t_1' из узла A' в узел B' идет ветвь с некоторой пометкой тогда и только тогда, когда в дереве t_1 из узла A в узел B идет ветвь с той же пометкой и узел A' удовлетворяет требованиям в узле A, и узел B' удовлетворяет требованиям в узле B.

Теорема 12. Задача установления функционального соответствия W/V-деревьев $T_{\chi^1}^{W/V}$ и $T_{\chi^2}^{W/V}$ принадлежит классу P комбинаторных задач с временной оценкой n^D , где $n = \max(|W_{\chi^1}^{W/V}|, |W_{\chi^2}^{W/V}|)$, $D = \sum_{i=1}^k \varphi(a_i)$, где

$\varphi = \begin{pmatrix} a_1, a_2, \dots, a_k \\ n_1, n_2, \dots, n_k \end{pmatrix}$ - матрица ограничений на характер

ветвления и на размещение пометок на ветвях из V, где $\{n_1, n_2, \dots, n_k\} \subset N$ - подмножество натуральных чисел.

Теорема 13. Задача поиска для каждого дерева $T_{\chi^l}^{W/V}$ из множества W/V-деревьев *forest1* дерева $T_{\chi^m}^{W/V}$ из множества *forest2*, функционально соответствующего $T_{\chi^l}^{W/V}$, принадлежит классу P комбинаторных задач с временной оценкой n^D , где множества *forest1* и *forest2* представимы в виде лесов ($forest1 = \langle V^{f1}, E^{f1} \rangle$, $forest2 = \langle V^{f2}, E^{f2} \rangle$) и $n = \max(|V^{f1}|, |V^{f2}|)$, а $D = \sum_{i=1}^k \varphi(a_i)$.

Служебная информация правил и относительность синонимических замен

Формат списка применимости правил к анализируемому дереву :

$$\left(\begin{array}{l} (\text{правило}(i) \text{ count}(i) C_0(i, \text{count}(i))) \dots \\ (\text{правило}(k) \text{ count}(k) C_0(k, \text{count}(k))) \end{array} \right), \text{ где}$$

$\text{count}(i)$ и $\text{count}(k)$ – счетчики вхождений в анализируемое дерево заменяемых i -м и k -м правилами поддеревьев;

$C_0(i, \text{count}(i))$, $C_0(k, \text{count}(k))$ – ключевые слова ЛСК.

Описание поля композиционных меток узла ГСС :

$$\left(\begin{array}{l} (\text{composition_label}(i, j, \text{count}(i)) \text{ count}(i) \text{ правило}(i)) \dots \\ (\text{composition_label}(k, l, \text{count}(k)) \text{ count}(k) \text{ правило}(k)) \end{array} \right),$$

где j, l – номера композиционных меток, соответственно, $\text{count}(i)$ и $\text{count}(k)$ -го вхождений заменяемых i -м и k -м правилами поддеревьев.

Формирование ЛФ-синонимических множеств

$\Pi^R \setminus \Pi_U^R$ - множество правил произвольных элементарных преобразований Δ -грамматики :

$$(\Pi^R \setminus \Pi_U^R) = \Pi_{LS}^R \cup \Pi_S^R, \quad \Pi_{LS}^R \cap \Pi_S^R = \emptyset, \text{ где}$$

Π_{LS}^R - множество лексико-синтаксических преобразований,

причем $\Pi_L^R \subset \Pi_{LS}^R$ - множество лексических преобразований без синтаксических замен;

Π_S^R - множество синтаксических преобразований без вспомогательных лексических замен.

$$\Pi_{LS}^R = \Pi_{LS_LCK}^R \cup \Pi_{LS_CONV}^R \cup \Pi_{LS_IMP}^R, \text{ где}$$

$\Pi_{LS_LCK}^R$ - множество двусторонних правил с сохранением валентных мест ЛСК;

$\Pi_{LS_CONV}^R$ - множество конверсивных замен с утратой валентности;

$\Pi_{LS_IMP}^R$ - множество смысловых импликаций;

$$\Pi_{LS_LCK}^R \cap \Pi_{LS_CONV}^R \cap \Pi_{LS_IMP}^R = \emptyset$$

$\Phi_i \Big|_{i=1}^n$ - последовательность ГСС :

$$\{\Phi_i\} = \{\Phi_i^{LCK}\} \cup \{\Phi_i^{NO_LCK}\}, \text{ где}$$

$\{\Phi_i^{LCK}\}$ - множество ГСС, к которым применимы правила множества $\Pi_{LS_LCK}^R$;

$\{\Phi_i^{NO_LCK}\}$ - множество ГСС, к которым не применимы правила множества $\Pi_{LS_LCK}^R$, причем

$$\{\Phi_i^{1_LCK}\} \cap \{\Phi_i^{1_NO_LCK}\} = \emptyset.$$

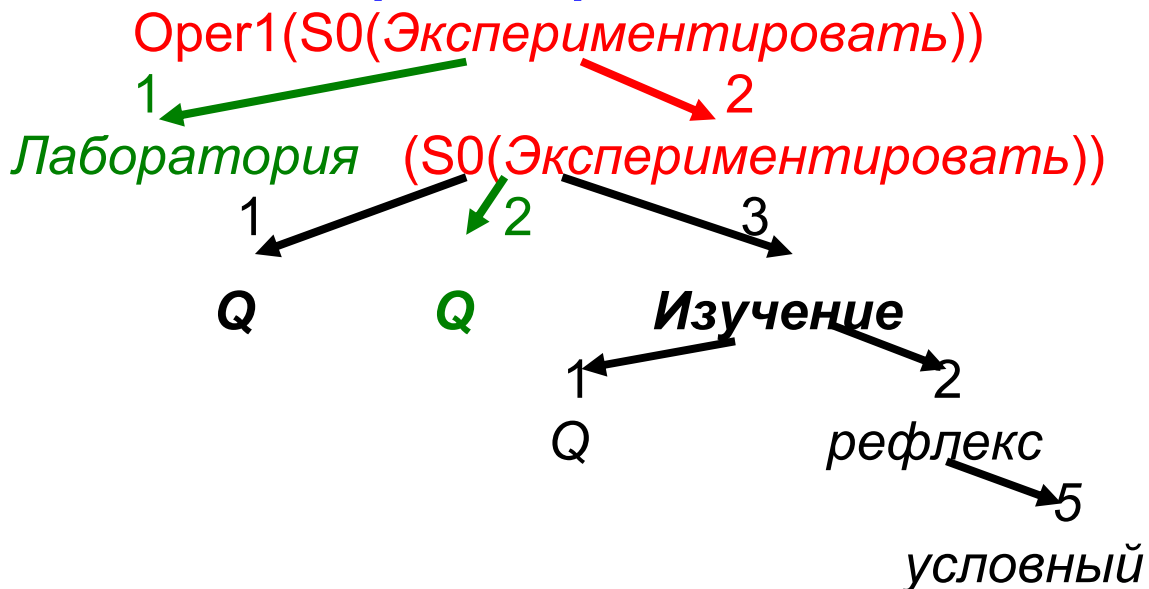
Таким образом, $\{\Phi\}_i = \{\Phi\}_i^{LCK} \cup \{\Phi\}_i^{CONV} \cup \{\Phi\}_i^{IMP} \cup \{\Phi\}_i^{SYNT}$,

причем $\{\Phi\}_i^{LCK} \cap \{\Phi\}_i^{CONV} \cap \{\Phi\}_i^{IMP} \cap \{\Phi\}_i^{SYNT} = \emptyset$

Пример построения образа суммарного смысла (этап 1, исходные деревья первого и второго предложений)

Первое предложение : *Лаборатория провела эксперименты по изучению условных рефлексов.*

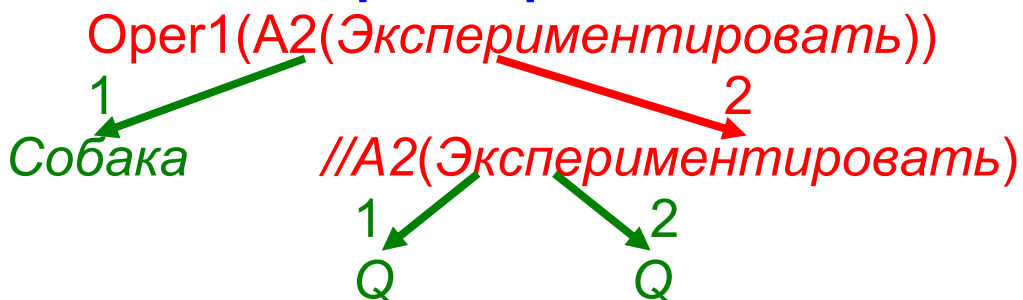
ГСС первого предложения :



Список применимости : ((17 1 Экспериментировать))

Второе предложение : *Подопытными животными были собаки.*

ГСС второго предложения :



Список применимости : ((16 1 Экспериментировать))

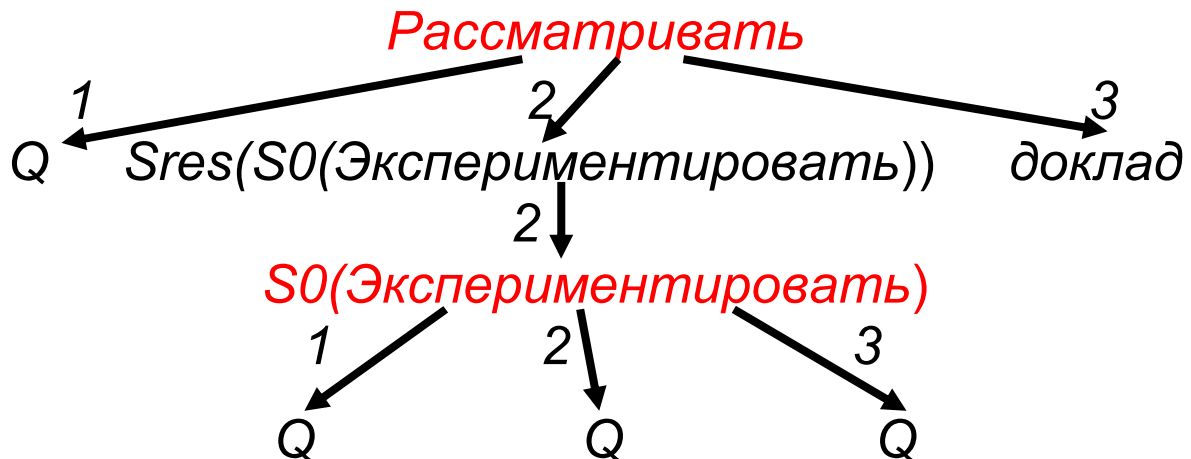
Обозначения :

- Поддерево, заменяемое лексическим правилом
- Поддерево, заменяемое синтаксическим правилом

**Пример построения образа суммарного смысла
(этап 2, исходные деревья третьего и четвертого
предложений)**

Третье предложение : Результаты экспериментов рассматривались в докладе на конференции.

ГСС третьего предложения :

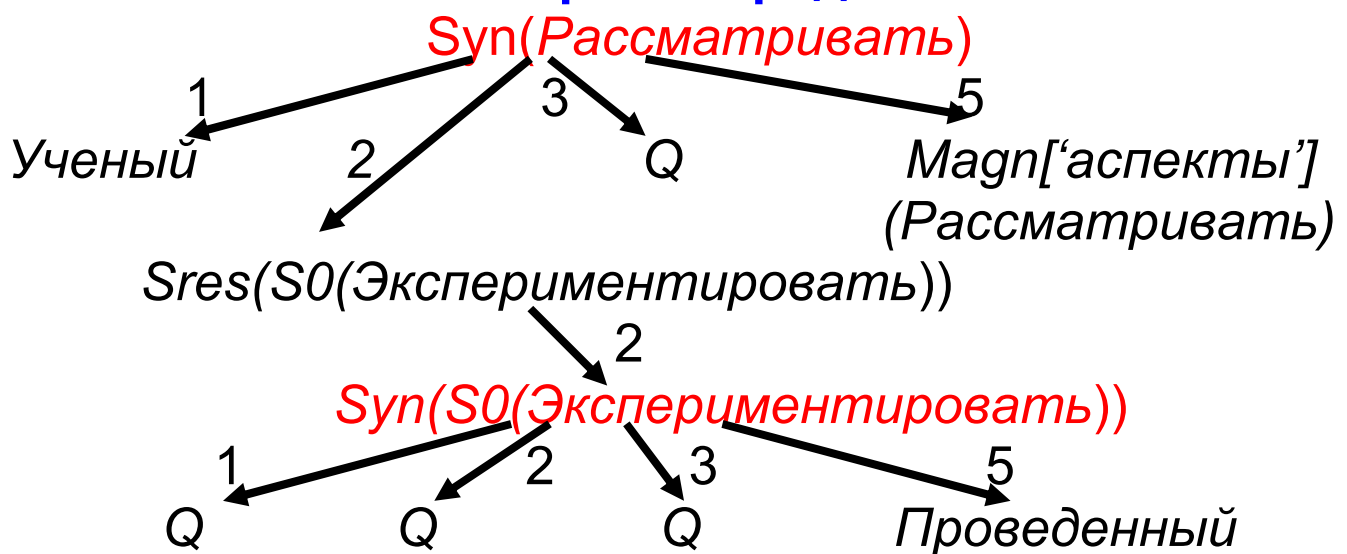


Список применимости :

((1 1 Рассматривать)(1 2 Экспериментировать))

Четвертое предложение : Ученый детально анализировал результаты проведенных опытов.

ГСС четвертого предложения :

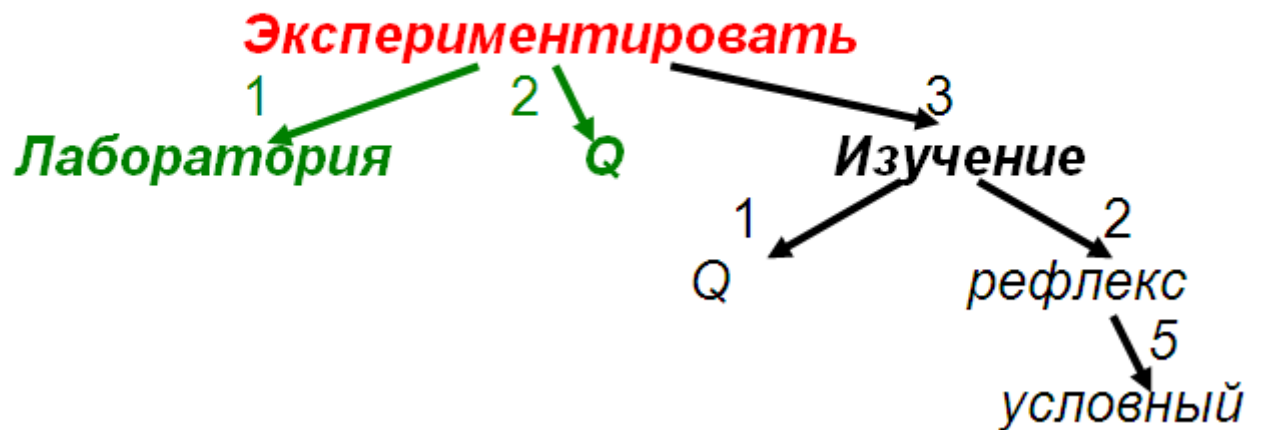


Список применимости :

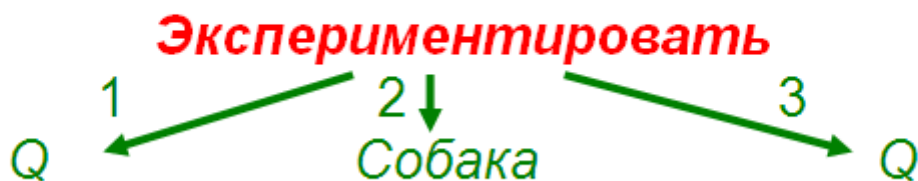
((1 1 Рассматривать)(1 2 Экспериментировать))

Пример построения образа суммарного смысла
(этап 3, преобразованные деревья первого и
второго предложений)

Преобразованная ГСС первого предложения :

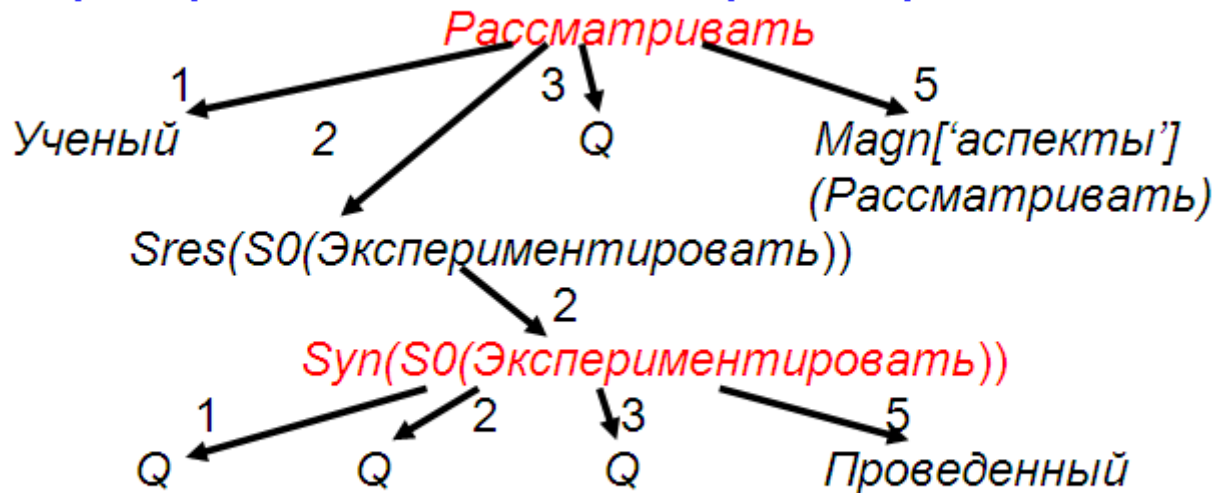


Преобразованная ГСС второго предложения :



**Пример построения образа суммарного смысла
(этап 4, преобразованное дерево четвертого
предложения и суммарная ГСС для первого и
второго предложений)**

Преобразованная ГСС четвертого предложения :



Список применимости : ((1 2 Экспериментировать))

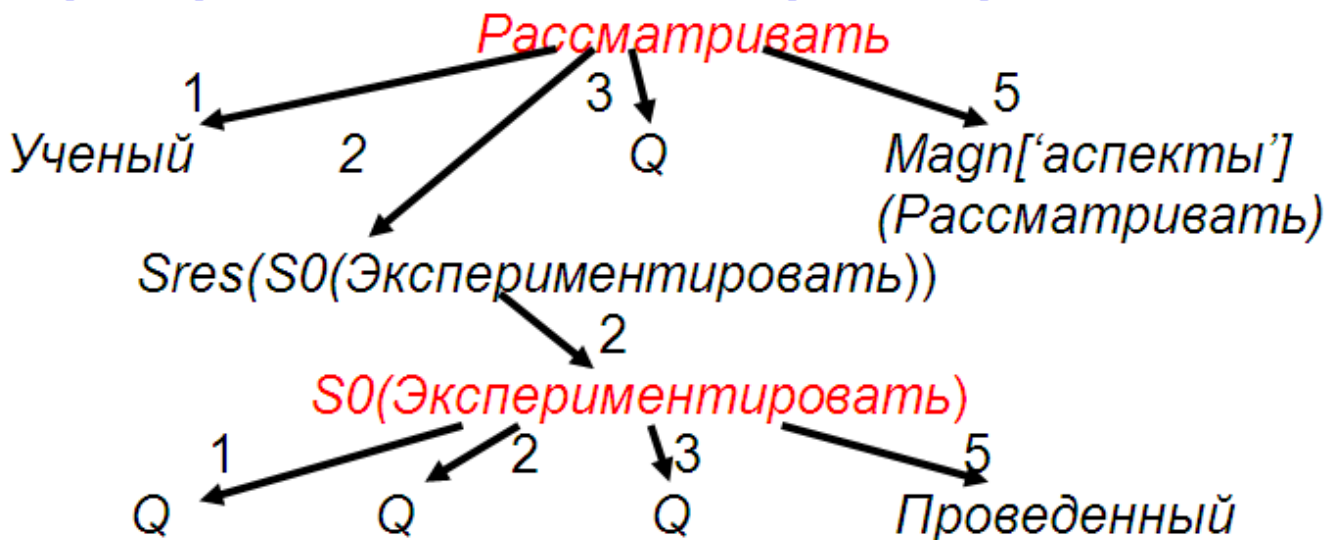
ГСС суммы первого и второго предложений



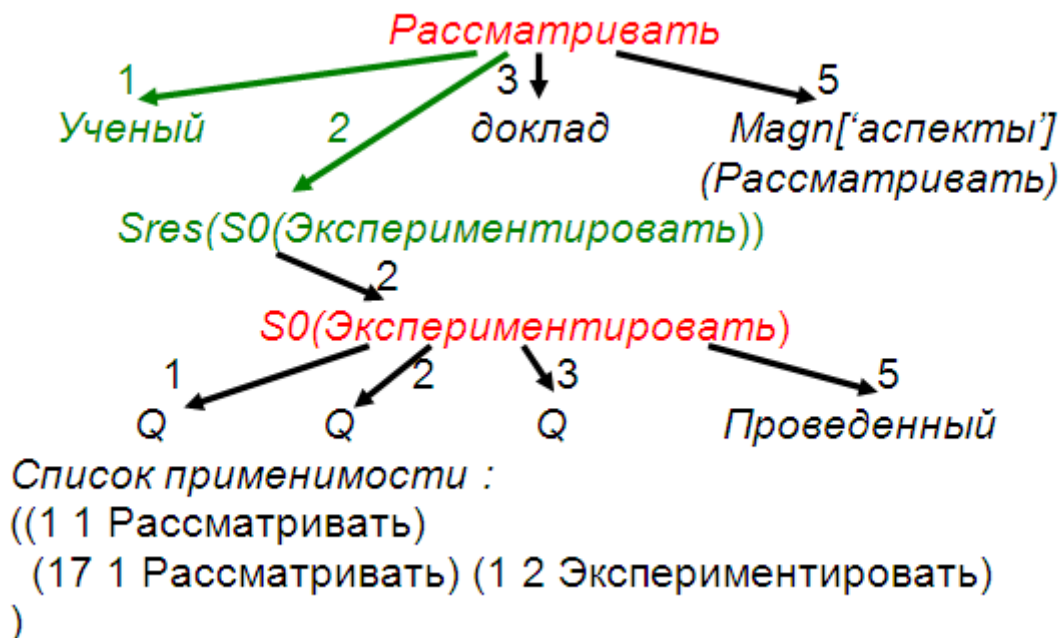
Список применимости : ((17 1 Экспериментировать))

Пример построения образа суммарного смысла
(этап 5, преобразованное дерево четвертого
предложения и суммарная ГСС для третьего и
четвертого предложений)

Преобразованная ГСС четвертого предложения :



ГСС суммы третьего и четвёртого предложений



Преимущества подхода :

- 1. Выделение сверхфразовых единств без существенного ограничения жанра анализируемого текста;**
- 2. Единые с задачей установления семантической эквивалентности механизмы оперирования лингвистическими знаниями.**