А.В.Сиренко

Методика интеграции базы данных лингвокультурного тезауруса с лексикографическими базами данных синонимов, антонимов[1]

Система когнитивных экспериментов использует ряд лексикографических объектов: ассоциативно-вербальную сеть, лингвокультурный тезаурус, орфографические словари, словари синонимичных отношений. В основе вычислительной модели лежит взвешенная контекстно-зависимая грамматика, образуемая последовательной интеграцией в нее указанных лексикографических объектов.

1. Структура ассоциативно-вербальной сети.

Ассоциативно-вербальная сеть является удобной моделью свободного ассоциативного эксперимента, обладающей свойством наглядности, широко проработанным математическим аппаратом.
Допущения ассоциативного эксперимента, позволяющие составить множества узлов и ребер ассоциативно-вербальной сети:

1. Стимул предъявляется текстовом виде;
2. Реакция фиксируется в текстовом виде;
3. Элементы с одинаковым написанием считаются эквивалентными.

При формировании ассоциативно-вербальной сети результаты опросов отдельных респондентов объединяются с учетом эквивалентности стимульно-реактивных узлов. При этом происходит суммирование индивидуальных ассоциаций респондентов в ассоциации усредненного носителя языка с увеличением значения весового компонента Ex события E_{StRk}.
Генеральная совокупность событий ассоциативного эксперимента представляет собой ассоциации между возможными парами стимул-реакция. В процессе эксперимента выборка представляет собой проявленные в опыте ассоциации, формирующие ассоциативные поля стимулов. В результате агрегации экспериментальных данных получаем квантитативные характеристики ассоциаций, которые могут быть преобразованы в статистическую вероятность Prob, с которой осуществлялся переход от стимула S к реакции R в эксперименте:

$$Prob(\mathrm{E}_{i,j}) =$$

$$\begin{cases} 0, \nexists E_m \in Sample: & E_m[St] = St_i \cap E_m[Rk] = Rk_j \\ \dfrac{E_m[Ex]}{\sum_{k=0}^{k<n} E_k}, \exists E_m \in Sample: & E_m[St] = St_i \cap E_m[Rk] = Rk_j \end{cases} \quad (1)$$

Выборка дает возможность приближенной оценки генеральной совокупности, при этом необходимо представить:

1. вероятностную оценку зафиксированных в эксперименте ассоциаций;

[1] В данной статье приводятся результаты исследований, выполненных при поддержке гранта РГНФ №12-04-12039в

2. оценку вероятности ассоциации, которая не была зафиксирована в эксперименте (распределение части вероятностного пространства по ребрам, имеющим нулевой вес/бесконечную длину).

К решению первой задачи применяются два подхода: определение вероятностных характеристик ассоциативного поля – вероятностей реакций, доверительных вероятностей, либо рассмотрение энтропии стимула в качестве характеристики меры снятия неопределенности о последующей реакции. Подобный разносторонний подход позволяет оценивать полноту ассоциативного эксперимента, а также вводить понятие ассоциативной силы слова (Навалихина, 2010).
Обобщенной моделью исходных данных для анализа «пространственно-временной структурности» и текста, и АВС является матрица связанности ЯЕ, элементами которой являются значения функции «силы связи»:

$$f_{ij} = F(\text{ЯЕ}_i, \text{ЯЕ}_j) \qquad (2)$$

Функция F интерпретируется в соответствии с некоторыми априорно заданными правилами. Для текста это может быть дистрибутивно-частотная связность языковых единиц.

2. Построение вычислительной модели когнайзера.

При рассмотрении данной методики считаем ассоциативно-вербальную сеть построенной, с рассчитанными частотными характеристиками, представляющую собой марковскую сеть ассоциативного эксперимента. Предобработка результатов когнитивного эксперимента заключается во вводе данных в реляционную систему управления базами данных программного комплекса.
Методика интеграции психолингвистических экспериментов предполагает:

1. Лемматизацию АВС и элементов тезауруса;
2. Построение взвешенной КС-грамматики на основе АВС;
3. Построение взвешенной КЗ-грамматики, включающей КС-грамматику и КЗ-правила на основе когнитивных единиц;
1. 4. Дополнение взвешенной КЗ-грамматики лексикографическими объектами (словарями).

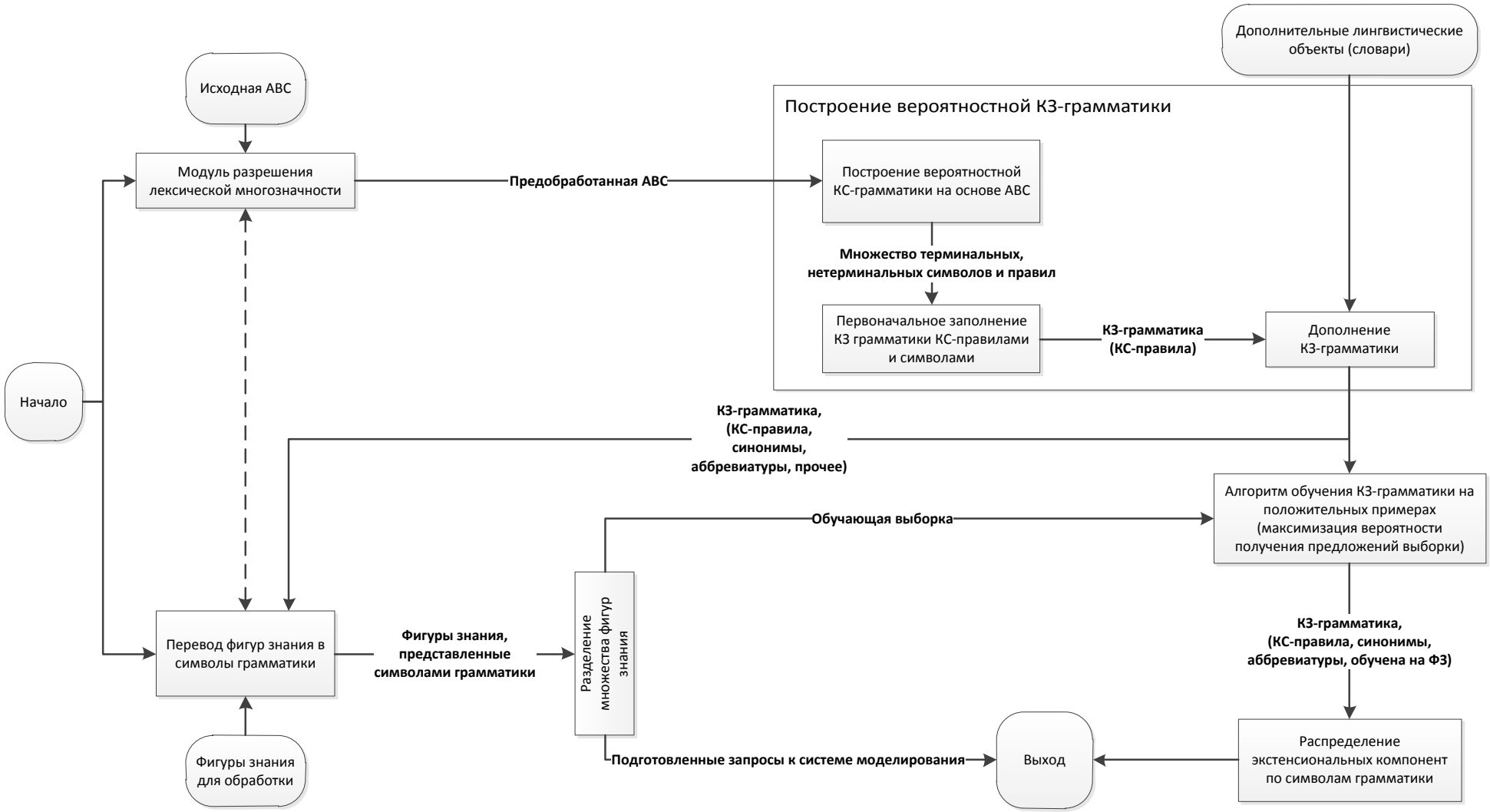

Рисунок 1. Интеграция исходных лексикографических объектов в грамматику

a. Методика лемматизации АВС и лингвокультурного тезауруса.

Автоматизированная система научных исследований проектировалась для автоматической обработки когнем тезауруса, что требует приведения более 16 000 знаков и формул смысла к множеству узлов ассоциативной сети. Хотя ассоциативная сеть, как представлено в работе (Караулов, 1993), содержит практически всю грамматику русского языка, грамматические формы и отношения могут присутствовать несколькими примерами. То есть не все многообразие словоформ, присутствующих в фигурах знания и потенциально присутствующих в запросах пользователей, есть в ассоциативной сети. Для увеличения доли автоматически обрабатываемых запросов производится лемматизация АВС и входящих запросов.

Другой причиной поддержки лемматизации в программном комплексе является явление омонимии. Омонимия может быть частичной либо полной. Для полных омонимов при одинаковом написании всех словоформ их смысловая нагрузка отличается (ключ, замок, лук). Частичные омонимы обладают различными наборами словоформ, имеющих общие элементы.

Стоит заметить, что омонимия не только препятствует лемматизации сети и входящих запросов, но и вносит ошибочные связи в ассоциативную сеть, представляя семантически разные узлы общим, объединяя их связи.

Цели методики лемматизации: формирование лемматизированной АВС, формирование таблиц лемматизации (множеств лемм и словоформ).

Лемматизированная АВС:

$$NET_{Lem} = \{CONS_{Lem}, LEMMS, WFS\},$$

где $CONS_{Lem}$ – связи лемматизированной сети, $LEMMS$ – множество лемм, WFS – множество словоформ.

$$CONS_{Lem} = \{conid_{Lem}, stid_{Lem}, rkid_{Lem}, fr_{Lem}, pr_{Lem}\},$$

где $conid_{Lem}$ – идентификатор связи лемматизированной сети, $stid_{Lem}$ – идентификатор узла-стимула, $rkid_{Lem}$ – идентификатор узла-стимула, fr_{Lem} – число фиксаций связи в эксперименте, pr_{Lem}- частотная характеристика связи.

$$LEMMS = \{lemid, lemname, lemrel\},$$

где **lemid** – идентификатор леммы, **lemname** – текстовое наименование леммы, **lemrel** – релятор леммы (в случае отсутствия – пустая строка). Термин релятор расшифровывается ниже.

$$WFS = \{wfid, wflem, wfname\},$$

где **wfid** – идентификатор словоформы, идентификатор соответствующей словоформе леммы, **wfname** – текстовое наименование словоформы.

Указанные множества обладают зависимостями:

$$lemma, stid_{Lem}, rkid_{Lem} \in \{lemid\}$$

Релятор – символ или слово, используемое для различения значений многозначного слова (ГОСТ 7.74-96, 1996). Релятор введен как атрибут леммы для различения полных омонимов. Таким образом, информационная технология позволяет различать полные омонимы, но, в силу трудоемкости разделения (десятки тысяч полных омонимов в русском языке), данная работа не была произведена.

Методика лемматизации включает этапы:

Этап 1. Генерация лемм и словоформ.

Лемматизация проведена с использованием орфографического словаря iSpell, используемого для проверки орфографии в среде Unix. Он включает в себя материалы орфографических словарей:

- Электронный орфографический словарь «Корректор» (120 тыс. словоформ);
- Грамматический словарь русского языка: словоизменение. Зализняк А.А. (110- тыс. словоформ);
- Сводный словарь современной русской лексики. Издательство Русский язык, 1991г. (170 тыс. словоформ).
- Русский орфографический словарь. Под редакцией В. Лопатина (180 тыс. словоформ).
- Прочие источники.

Выполнение **1-го этапа** методики лемматизации формирует: лемм - 127 000, словоформ - 1 300 000.
Правила генерации словаря iSpell:

$$ISPELL = \{BASES, AFFRULES\},$$

где $BASES$ – леммы словаря, $AFFRULES$ – правила преобразования.

$$BASES = \{base, < flag >\},$$

где **base** – текстовое наименование леммы, <**flag**> - множество идентификаторов правил генерации.

$$AFFRULES = \{ruleid, prev, notprev, oldp, cutp, addp\},$$

где **ruleid** – идентификатор правил преобразования леммы, **oldp** – требуемое окончание леммы, **prev** – символьная последовательность, предшествующая **oldp**, **notprev** – символьная последовательность, которой не должно быть перед **oldp**, **cutp** – отсекаемое от леммы окончания, **addp** – добавляемое к лемме окончание после отсечения **cutp.**

$$flag \in \{ruleid\}$$

Этап 2: Обработка дефективных данных.

В результате генерации лемм и словоформ во множестве лемм оказывается часть элементов, не имеющих собственных словоформ и относящихся к другим леммам в качестве словоформы, и также лемм, которые должны быть объединены. Это выполняется применением ряда выделенных вручную правил преобразования.

Этап 3: Назначение лемм узлам АВС.

Для каждого узла исходной ассоциативной сети производится поиск соответствующей ему словоформы, после чего – леммы. Этим переходом формируется множество связей лемматизированной сети $CONS_{Lem}$.В силу омонимии, часть связей на данном этапе не имеет назначенных лемм.

Этап 4: Обработка частичных омонимов.

Под частичной омонимией подразумевается совпадение отдельных словоформ у разных по написанию и смыслу лемм. Примерами частичных омонимов могут быть: чеки (чек, чека), белок (белок, белка), шерсти (шерсть, шерстить). Анализ ассоциативной сети

показал, что она содержит 1030 частичных омонимов, что делает неопределенными для лемматизации 25 880 связей.

1. Обработка неоднозначных связей АВС;
2. Создание омонимичных лемм.

Для поддержки решения поставленных в 4-м этапе задач, программный комплекс осуществляет взаимодействие с экспертом посредством экранных форм. Связи ассоциативной сети, распределяются среди «конкурирующих» лемм экспертным мнением. Решение сохраняется в виде текстовых файлов (скриптов), позволяющих повторить преобразования при повторной обработке данных и проверить изменения после работы. Формат файла представлен ниже.

```
СЛОВОФОРМА

# ЛЕММА1->РЕЛЯТОР_ЛЕММЫ1

СЛОВОФОРМА_СТИМУЛ->СЛОВОФОРМА_РЕАКЦИЯ->ЧИСЛО_СВЯЗЕЙ

...

# ЛЕММА2->РЕЛЯТОР_ЛЕММЫ2

СЛОВОФОРМА_СТИМУЛ->СЛОВОФОРМА_РЕАКЦИЯ->ЧИСЛО_СВЯЗЕЙ

...
```

Рисунок 2. Формат скрипта обработки омонимии АВС

```
шерсти

# шерстить->

# шерсть->

комок->шерсти->1

клубок->шерсти->7
```

Рисунок 1. Пример скрипта обработки омонимии АВС

В результате этапов лемматизации, итоговая ассоциативно-вербальная сеть является лемматизированной в соответствии со словарем словоформ.

Число	Исходная сеть	Лемматизированная сеть
Узлов	103 000	63 700
Связей	457 000	394 000
Стимулов	6665	3833

Таблица 1 Сокращение размерности сети при лемматизации

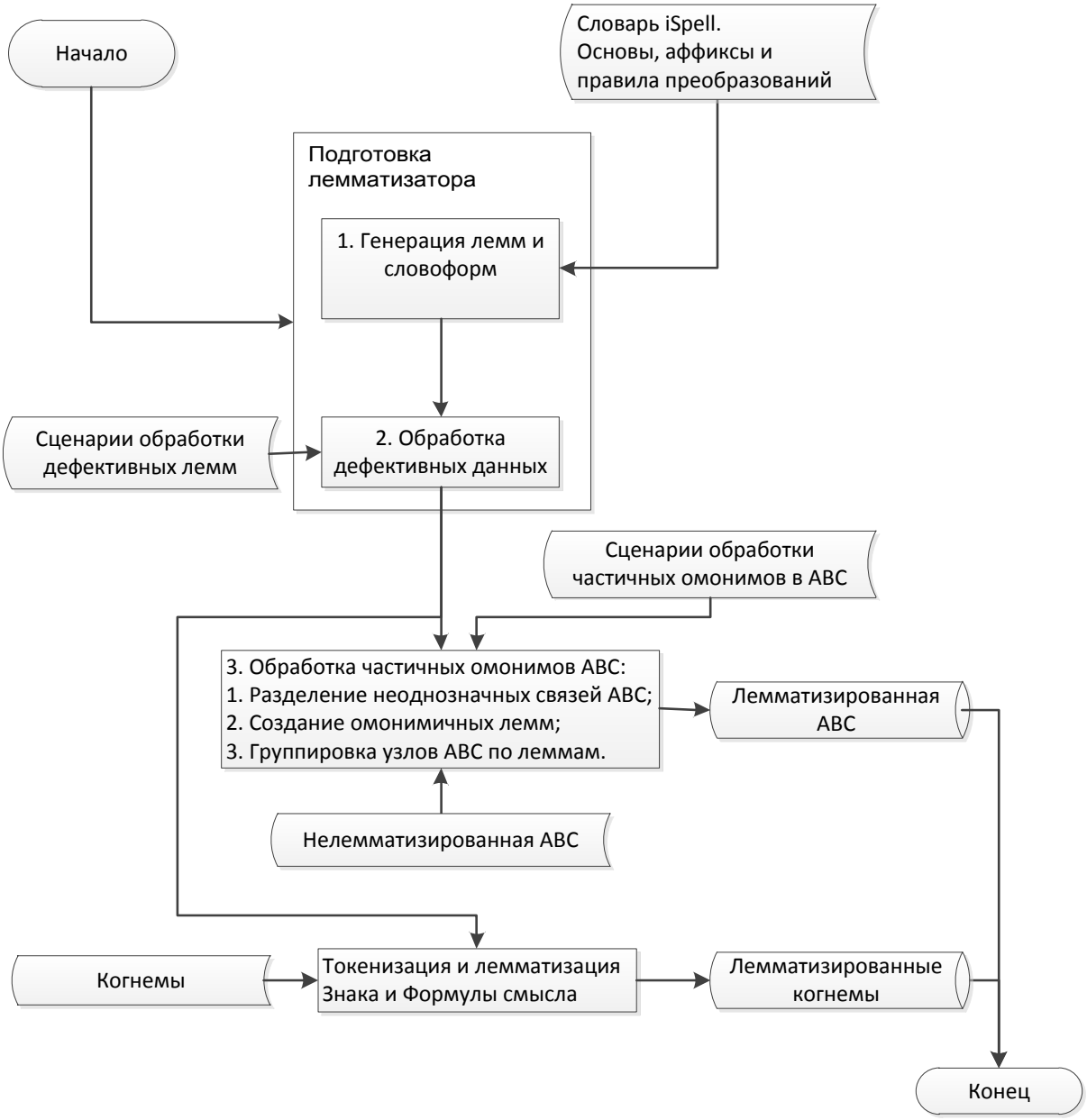

Рисунок 4. Методика лемматизации.

2.5. Методика построения взвешенной контекстно-свободной грамматики на основе ассоциативно-вербальной сети.

Ассоциативно-вербальная сеть позволяет построить событийно-статистическую модель вербального сознания, основанную на ассоциативных цепочках. Данная модель имеет своим недостатком избыточность результатов моделирования (получаемых цепочек смены активных мыслезнаков), возникающих в силу:

- Логической неадекватности сети [Джексон П., 2001];
- Сильной связности;
- Недостаточного присутствия лексической и экстралингвистической информации в модели;
- Потери синтаксической информации запроса при переводе его в ассоциативно-вербальную сеть.

На практике, это приводит к необходимости либо ручного отбора цепочек исследователем, либо их ручным построением, с использованием программных инструментальных средств для отображения ассоциативных полей.
Достоинством использования вероятностной модели АВС (представления в виде Марковского процесса) является простота построения, наличие количественной (частотной) оценки результатов, известные алгоритмы работы с подобными моделями.
Недостатки АВС:

- Не предполагает наличия контекста, как правило, осуществляется переход к марковским цепям (Manning, et al., 1999).
- Плохо отражает рекурсивные свойства языка.
- Сложность интеграции дополнительной информации в модель. При этом используется преобразование исходного запроса, а во время моделирования используется только АВС .

Недостатки могут быть устранены построением вероятностной грамматики на основе эмпирических данных ассоциативного эксперимента и последующем ее расширении.

Методика построения грамматики.

Согласно определению, грамматикой является тройка вида:

$$\boldsymbol{G} = (\boldsymbol{T}, \boldsymbol{N}, \ \boldsymbol{P}), \qquad (3)$$

где T - множество терминальных символов грамматики,
N – множество нетерминальных символов грамматики,
P – множество правил грамматики.
Обозначим объединенным алфавитом грамматики

$$\mathbf{V} = \boldsymbol{T} \cup \boldsymbol{N} \qquad (4)$$

$$v_i = \ (t_i \vee n_i);$$

1. Формируем множества символов грамматики: $\boldsymbol{N} = \{\boldsymbol{S} \mid \boldsymbol{SR}\}, \boldsymbol{V} = \{\boldsymbol{R}\}$;
2. Формируем правила грамматики на основе правил перехода в ассоциативной сети:

$$R = (ls, rs,\ pr), \qquad (5)$$

где ls $= \{v_i\} : ls \cap N \neq \emptyset$ – левая часть правила;
rs $= \{v_i\}$ – правая часть правила;
$pr = E_{ij}(prob)$ – вероятность фиксации ассоциации между узлами i и j в процессе свободного ассоциативного эксперимента при стимуле j;

3. Добавляем правила эквивалентности: словари синонимов, аббревиатур, устойчивых выражений, допускающих только прямое совпадение. Полученная грамматика теряет стохастические свойства, становится взвешенной. В терминах вероятностной грамматики правила эквивалентных замен имеют условно единичную вероятность, то есть служат для объединения символов грамматики и увеличение числа нетерминальных символов. На примере синонимичных отношений:

   $RSym_{ij} = \{ S_i, S_j, pr = 1 : i \neq j,\ S_i, S_j \in Sinset_k \},$
   где $Sinset_k$ – синонимичная группа;
   При построении подобных правил для всех элементов синонимичного ряда обеспечивается возможность эквивалентной замены символов внутри группы.
4. Добавляем терминальные символы, соответствующие нетерминальным и правила, позволяющие завершить выв в любой момент (заменив все нетерминальные символы терминальными). Этот шаг является формальной частью методики. Вычислительная процедура завершает работу в требуемый момент без замены симловов, считая каждый символ в том числе терминальным.

Построенную таким образом грамматику можно классифицировать как грамматику общего вида, то есть не имеющую ограничений на правила (Белоусов, и др., 2004). Для формального соответствия определению в состав грамматики неявно добавляется начальный символ S, из которого с помощью правил можно перейти в каждый нетерминальный символ $n_i \in N$.
Этапы добавления в грамматику синонимичных правил представлены на иллюстрации ниже (Рисунок). Согласно схеме, обработка естественно-языковой пропозиции лемматизатором предполагает его разделение на составляющие (токенизацию), проверку наличие словоформ в лемматизаторе. Если словоформ нет, токены пополняют базу лемматизатора. Порядок токенов в пропозиции сохраняется.

Рисунок 5. Введение отношений эквивалентности в грамматику.

2.6. Расширение грамматики экстенсиональными компонентами когнем.

Экстенсиональными компонентами фигур знания являются СПОСОБ (**method**), ФУНКЦИЯ (**func**) и ОБЛАСТЬ (**AREAS$_i$**).
СПОСОБ определяется характером связи между формулой смысла и знаком, т.е. эта компонента индивидуальна для фигуры знания и не является характеристикой мыслезнака, а значит и символа грамматики.
ФУНКЦИЯ отражает важность мыслезнака для носителя языка и обладает тем же свойсвом принадлежности полной когнитивной единице, что и СПОСОБ, например:

```
   [Груша] – [Из какого плодового дерева изготавливают музыкальные
инструменты?] – [Ретушь].
```

Таким образом, использование СПОСОБА и ФУНКЦИИ не может применяться непосредственно для разметки символов грамматики.
В результате выполнения методики построения взвешенной КС-грамматики на основе АВС формируется КС- грамматика, содержащая правила:

- Ассоциативные правила на основе АВС;
- Правила синонимичных отношений;

Большинство человеко-ориентированных коллекций декларативных знаний являются либо неструктурированными (ЕЯ описание), либо имеют вид фреймов, содержащих ЕЯ пропозиции в своем составе. Согласно методике САЭ, условия формирования пары стимул-реакция вели к минимальной структуре ассоциативной пары. АВС содержит единичные лексические единицы, что приводит к формированию контекстно-свободных

правил грамматики. Словосочетание в анкете ассоциативного эксперимента является отступлением от методики эксперимента и расценивается экспериментаторами как брак. Ассоциативно-вербальная сеть для каждого концепта содержит: денотат концепта, неупорядоченное множество взвешенных направленных связей к другим концептам. Хотя ЕЯ-описание также определяет концепт через другие, но способом «упаковки» концептов в систему с помощью синтаксической структуры. При переходе на уровень семантики синтаксическая структура преобразуется в семантическую. Не рассматривая в данной работе проблематику соотношения синтаксиса и семантики отметим, что ЕЯ-описание формирует семантическое содержание, которое:

- Не совпадает с каким-либо концептом из совокупности входящих в ЕЯ-описание, по меньшей мере, специфицируя или расширяя его.
- Зависит от базы знаний воспринимающего описание субъекта / информационной системы.

Целью интеграции когнитивных единиц в ассоциативную грамматику является создание контекстно-зависимых правил и правил общего вида, позволяющих сделать вывод в грамматике более специфицированным, учитывающим декларативные знания о ЯКМ.
Задачи, которые необходимо решить:

1. Формирование контекстно-зависимой левой части правил на основе ЕЯ-описания концепта;
2. Определение квантитативной характеристики созданного КЗ-правила;
3. Применение правила при неполном совпадении обрабатываемой пропозиции и правила.

Вследствие динамичности естественного языка, лексикографические объекты и базы знаний на их основе обладают свойствами неполноты и открытости – элементы базы знаний не могут быть описаны тем же множеством элементов. Построенная грамматика принципиально не может покрывать многообразие лексики возможных запросов. Чтобы не игнорировать «новую» лексику при обработке, на основе вероятностной грамматики G будет определена расширенная грамматика. Под расширением мы понимаем способность грамматики дополнять множество символов в произвольный момент времени, в том числе в момент поступления запроса.
В методике лемматизации приведен порядок преобразования элементов когнем ЗНАК и ФОРМУЛА_СМЫСЛА в множество лемматизированных токенов. На этапе интеграции когнем в грамматику дополнительно используется множество стоп-слов (StopWords), полученное подсчетом корпуса текстов различной тематики, отбором наиболее частотных и фильтрацией этого множества экспертом.

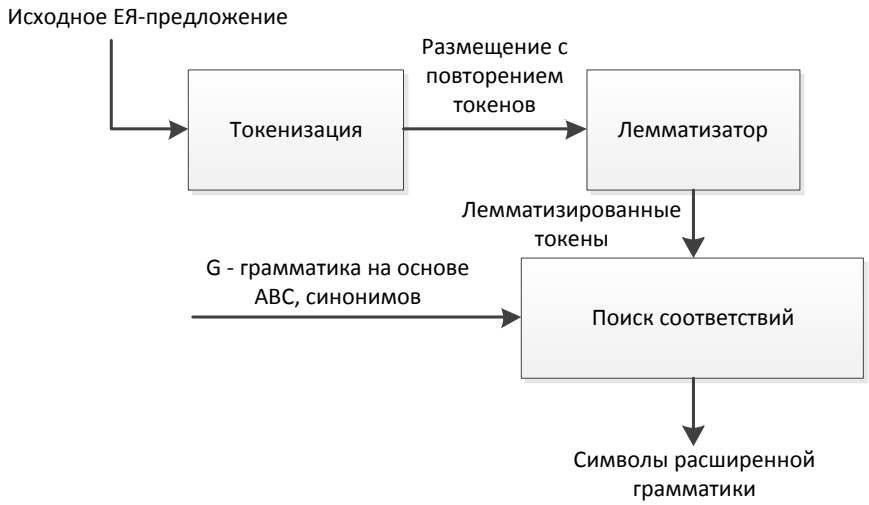

Рисунок 6. Преобразование ЕЯ предложения в символы расширенной грамматики.

Рисунок 7. Интеграция экспериментальных данных в КЗ-грамматику.

Обозначим расширенной грамматикой

$$GExt = (T_{Ext},\ N_{Ext}, P_{Ext},\ S_{Ext}), \qquad (6)$$

где $\boldsymbol{T_{Ext}}$ - множество терминальных символов,
$\boldsymbol{N_{Ext}}$ – множество нетерминальных символов,
$\boldsymbol{P_{Ext}}$ – множество правил,
$\boldsymbol{S_{Ext}}$ – начальный символ грамматики;
Символ расширенной грамматики:

$$SymExt = (id, name, symKey, SW), \quad)$$

где symKey – ключ символа грамматики G;
id – идентификатор;
name – текстовое представление символа;
SW – булевый признак принадлежности к стоп-словам.
Расширенная грамматика заполняется на основе взвешенной КС-грамматики G и результатов когнитивного эксперимента следующим образом:

1. $TExt$ и $NExt$ заполняются на основе соответствующих в G через тривиальную процедуру создания символов $GExt$;
2. $PExt$ заполняется на основе правил грамматики G;
3. Множество когнем преобразуется в правила расширенной грамматики через преобразование формулы смысла и знака в последовательности символов расширенной грамматики, соответствующие левой и правой части правила, соответственно;
4. $TExt$ и $NExt$ корректируются в соответствии с $PExt$.

Если знак или формула смысла когнемы после преобразования не содержат символов расширенной грамматики с установленным соответствующим символом из G и не принадлежащих к стоп-словам, дальнейшая работа с когнемой не производится.
Представим создание грамматики GExt в виде функции исходных данных и параметров:

$$GExt = f\left(G, G_{Threshold,}\, COGNEMS, AreaOption, MethodOption\right), \quad (8)$$

где G – ранее введенная КС-грамматика, COGNEMS – множество когнем,
$G_{Threshold}$ – пороговое значение квантитативного параметра правил из G для включения в GExt,
AreaOption – множество когнитивных областей, когнемы которых будут включены в GExt,
MethodOption – множество способов представления ФОРМУЛЫ_СМЫСЛА, которым должна соответствовать когнема для включения в GExt.

Литература.

Навалихина, 2010 Изменение энтропии в пространстве индивидуального поля. *Электронный научный журнал Курского государственного университета.* [В Интернете] 2010 г. http://tl-ic.kursksu.ru/pdf/007-15.pdf.

Караулов, Ю. Н. 1993. *Ассоциативная грамматика русского языка.* Москва : Русский язык, 1993. стр. 330. ISBN 5-200-01765-3.

ГОСТ 7.74-96. 1996. Информационно-поисковые языки. 1996 г.