или более (по умолчанию разделителем будем считать пробел). Для того чтобы иметь возможность перевести курсор на позицию элемента, необходимо, чтобы в структуре была представлена информация о местоположении каждого отдельного слова. Для всех узлов структуры, являющихся *конечными* или *определяющими*, создается отдельная таблица, в которой хранится информация о каждом элементе.

В предложенном случае необходимо хранить следующую информацию об элементе:

– принадлежность элемента к документу $\varepsilon \in P$, где $\varepsilon$ – индекс элемента (он же индекс таблицы), $P$ – индекс документа. Следует отметить, что $\varepsilon = \mu$, где $\mu$ – индекс узла. Индекс узла назначается всем *конечным* и *определяющим* узлам структуры, транзитным узлам индекс не назначается;

– индекс предыдущего и индекс следующего элемента в данном документе $\varepsilon_{i-1} < \varepsilon_i < \varepsilon_{i+1}$. Эти данные нам понадобятся для выполнения сложных запросов и осуществление лингвистического анализа текста;

– положение элемента $\varepsilon$ в документе $P$ в виде смещения относительно начала документа – обозначим его как $\theta$. По этому параметру, исходя из $\varepsilon \in P$, можно будет однозначно восстановить исходный документ из оптимизированной структуры. Таким образом, ограничение необратимости преобразования, о котором мы говорили выше – снимается.

Содержание расширенной базы напрямую зависит от требуемой функциональности системы. В случае дополнительных требований информативность расширенной базы может быть увеличена путем добавления необходимых полей в таблицу элемента (это может быть время создания документа, лингвистические характеристики: род, число, падеж и т. д.). Чем информативнее расширенная база, тем больше будет возможностей для проведения анализа, тем качественнее может быть работа всей поисковой системы в целом.

**Библиографические ссылки**

1. Талантов М. Поиск в Интернете: подводные камни // КомпьютерПресс. 1999. № 9. С. 46–52.

2. Мультилингвистическая модель распределенной системы на основе тезауруса / С. В. Рогов, П. В. Зеленков, И. В. Ковалев, М. В. Карасева // Вестник СибГАУ. 2008. Вып. 1 (18). С. 26–28.

3. Карцан И. Н., Лохмаков П. М., Цветков Ю. Д. Интеллектуализация поиска информации в корпоративных системах // Вестник НИИ СУВПТ. 2006. Вып. 23. С. 141–156.

N. A. Raspopin, M. V. Karasyova, P. V. Zelenkov, E. V. Kayukov, I. V. Kovalyov

**MODELS AND METHODS OF DATA COLLECTING AND PROCESSING**

*The paper considers the structure of data presentation, which meet certain requirements. The given additional data structure is developed, proceeding from requirements of the retrieval system and necessary system functionality.*

*Keywords: optimization, retrieval system, extended base, data collecting and processing.*



УДК 81`322

К. В. Сафонов, Д. В. Личаргин

**НЕКОТОРЫЕ ПРИНЦИПЫ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ УЧЕБНЫХ МАТЕРИАЛОВ НА ОСНОВЕ БАЗ ЗНАНИЙ И ЛИНГВИСТИЧЕСКОЙ КЛАССИФИКАЦИИ**

*Рассматриваются модели и средства генерации осмысленного подмножества естественного языка учебных курсов. В частности, задается семантическое понятийное пространство слов языка. Ставится цель построить модель генерации текстов для учебных курсов по английскому языку, формулируются задачи ее применения на основе порождающих грамматик над ориентированным лесом строк. Делается вывод о специфике и структуре модели генерации учебных курсов.*

*Ключевые слова: генерация естественного языка, семантические признаки, классификации слов и понятий языка, генерация учебных материалов.*

На современном этапе актуальной является проблема автоматизации систем письменного и устного перевода для различных языков, экспертных, поисковых систем и систем реферирования. Для решения данных задач успешно реализуются различные теории, концепции и программные системы. Многочисленные работы в области семантики, дискретной математики, лингвистики и искусственного интеллекта дают надежду на решение в ближайшем будущем многих проблем формализации естественного языка и прохождения теста Тьюринга во все более жестких для тестовых систем условиях. Особенно важной

оказывается проблема автоматической генерации учебных материалов как частного случая текстов на естественном языке.

**Уровень разработки.** Для решения проблемы генерации осмысленной речи на сегодня используется широкий инструментарий как семантики, так и искусственного интеллекта в рамках понятийного аппарата и различных моделей математической семантики. В частности, для анализа естественного языка традиционно используются следующие модели и средства: метод онтологий, метод лингвистической классификации, метод многомерного представления данных, OLAP-системы, реляционные базы данных, фреймы, инструментарий системного анализа. Также используются порождающие грамматики, в частности, порождающие грамматики Монтегю и грамматики сложения деревьев, семантические сети, теория графов и метод резолюций, гибридные системы, а также лингвистические методы, такие как компонентный анализ, валентностное представление слов языка, парадигматический метод, методы американского структурализма и др.

Новизна данной работы состоит в нахождении понятийного описания единиц естественного языка, способах задания и определения критериев осмысленности фраз на естественном, в частности, английском языке, а также нахождении некоторых принципов проецирования понятийного пространства слов языка на иерархическую структуру учебного курса, в частности по английскому языку.

Основная идея работы состоит в представлении единиц естественного языка в виде множества деревьев, составляющих ориентированный лес естественного языка (или, иначе, лес текста), которым соответствует многомерное пространство векторизованных данных. Необходимо задать иерархию деревьев и понятийное векторное описание для каждого из них. Далее понятийное пространство единиц языка проецируется на структуру электронного учебного курса. Способы проецирования – многовариантны.

Цель работы – построить модель генерации текстов для учебных курсов по английскому и принципиально любому другому языку. Задачи работы состоят в применении данной модели на основе порождающих грамматик над ориентированным лесом строк, использовании классификации семантических понятий и слов естественного языка, векторизованной на базе традиционных неукорачивающих порождающих грамматик, определение места этого метода в системе языка в целом.

**Модель языка.** Лингвистические системы трудны для моделирования. Тем не менее можно выделить четкую структуру теста на естественном языке. Текст состоит из иерархии ярусов, срезов системы естественного языка:

– множество бессмысленных текстов – $L(21)$;

– множество грамматически осмысленных текстов – $L(20)$;

– множество семантически осмысленных текстов – $L(19)$;

– множество всех существующих текстов – $L(18)$;

– библиотека – $L(17)$;

– классификация текстов в каталоге библиотеки – $L(16)$;

– серия книг – $L(15)$;

– набор томов – $L(14)$;

– том – $L(13)$;

– главы – $L(12)$;

– разделы/параграфы – $L(11)$;

– абзацы – $L(10)$;

– пары и цепочки предложений – $L(9)$;

– сложные предложения – $L(8)$;

– простые предложения – $L(7)$;

– конструкции – $L(6)$;

– синтагмы – $L(5)$;

– фразеологизмы – $L(4)$;

– словоформы – $L(3)$;

– морфемы – $L(2)$;

– буквы – $L(1)$;

– признак буквы – $L(0)$.

Текст естественного языка состоит из следующих срезов/аспектов:

– срез написания (цепочка букв – символов алфавита) – $S(0)$;

– срез произношения (цепочка звуков) – $S(1)$;

– грамматический срез (добавление грамматических конструкций и категорий) – $S(3)$;

– семантический срез (шаблоны подстановок смысловых единиц языка) – $S(4)$;

– текстологический срез (шаблоны заполнения относительно статической структуры текста) – $S(5)$;

– срез актуального членения предложения (тема, рема, модальность, пояснение и др.) – $S(6)$;

– стилистический срез (множества особенностей всех предыдущих срезов в зависимости от ситуации и манеры речи) – $S(7)$ и др. [1–6].

Система естественного языка состоит из следующих элементов: единиц естественного языка и связей между ними: синтагматических, контекстуальных, в частности грамматических, семантических и иных связей.

Система естественного языка может рассматриваться (рис. 1) как множество строк текстов, разделенных на отдельные сегменты. Элементы этой системы – единицы языка – находятся в пространствах состояний – в виде, в частности, классификаций единиц языка, например, слов.

Для каждого уровня и среза языка имеет место особое пространство состояний единиц языка определенного уровня.

Пространства состояний единиц языка могут быть представлены в виде классификации с различным упорядочением семантических признаков классификации и соответственно упорядочиваемым узлам классификации.

Единицы естественного языка включаются в классы единиц естественного языка, пересечение и комбинаторика которых порождает его парадигматическую систему – фрагменты реляционной базы данных.

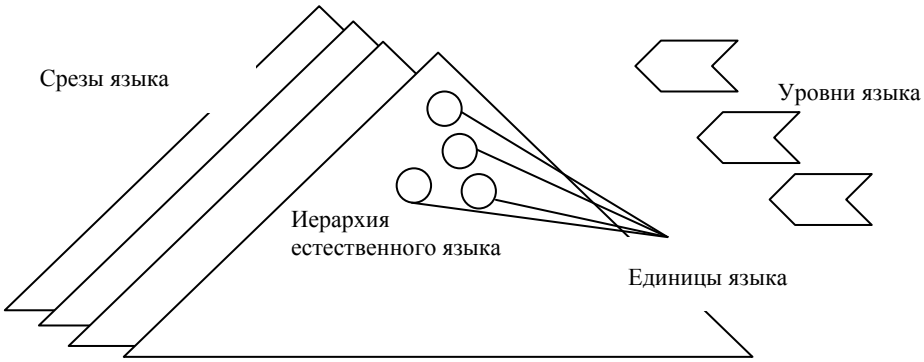


Рис. 1. Лес естественного языка как иерархическая система

Парадигмы естественного языка являются подмножествами многомерных пространств, или (что эквивалентно) древесных иерархий естественного языка, представляемых на основе векторов признаков древесной классификации, т. е. векторов координат многомерного пространства для состояний единиц естественного языка [1].

Приведем пример семантической классификации слов естественного языка. Последняя задается на основе вектора классификации, задаваемой порождающей грамматикой следующего вида.

На основе классификации сем естественного языка предлагается вектор классификации понятий естественного языка из пяти координат. Значения координат вектора $G = F[L(3),S(4),\ldots]$ задаются при помощи порождающих грамматик следующего вида.

1. Первый уровень классификации понятий, соответствуют признаку $G_1$ вектора $G$. Положим $G_1$ = {НЕЧТО, ОТНОШЕНИЕ, СОЗНАНИЕ, ИДЕЯ, ИНФОРМАЦИЯ, МЕСТО, ПРЕДМЕТ, СУЩЕСТВО}.

2. Второй уровень классификации понятий представлен признаком $G_2$. Множество $G_2$ значений признака классификации задается множеством правил порождающей грамматики: {$S \rightarrow Fd$, $S \rightarrow Fx$, $d \rightarrow$ЖИВОГО, $d \rightarrow$НЕЖИВОГО, $x \rightarrow$ КОТОРОГО ЖИВОЕ, $x \rightarrow$ КОТОРОГО НЕЖИВОЕ, $F \rightarrow$ЧАСТЬ (OF), $F \rightarrow$ВНУТРИ (IN), $F \rightarrow$НА ПОВЕРХНОСТИ (ON), $F \rightarrow$ОКОЛО (AT)}, где понятие ОКОЛО обозначает любое ненулевое расстояние между объектами.

3. Третий уровень классификации понятий определяется признаком $G_3$, $G_3$ = {*X–y* (*сущность*), *X–X–y* (*сущность чего-то*), ОТНОШЕНИЕ–*X–y* (*свойство*), ОТНОШЕНИЕ–*X–X–y* (*связь*), ОТНОШЕНИЕ–СУЩЕСТВО–*X–y* (*действие*), ОТНОШЕНИЕ–СУЩЕСТВО–*X–X–y* (*соединение*), ОТНОШЕНИЕ–СУЩЕСТВО–СУЩЕСТВО–*X–y* (*презентация*), ОТНОШЕНИЕ–СУЩЕСТВО–СУЩЕСТВО–*X–X–y* (*обмен*)}, где *X* – любая из основных сем, определенных на первом уровне классификации, *y* – любая последовательность таких сем. *X* выделяется как главная по смыслу сема. Знак «–» используется в данном случае для обозначения конкатенации. В круглых скобках приведены смысловые пояснения.

4. Множество $G_4$ значений признака $G$ задается множеством правил порождающей грамматики: {$S \rightarrow P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6 \cdot P_7 \cdot P_8$, $P_1 \rightarrow g \cdot$КОЛИЧЕСТВО, $P_1 \rightarrow \lambda$, $P_2 \rightarrow g \cdot$УСТОЙЧИВОСТЬ, $P_2 \rightarrow \lambda$, $P_3 \rightarrow g \cdot$ПОЗИТИВНОСТЬ, $P_3 \rightarrow \lambda$ , $P_4 \rightarrow g \cdot$СПЕКТР, $P_4 \rightarrow \lambda$, $P_5 \rightarrow g \cdot$ИНФОРМАТИВНОСТЬ, $P_5 \rightarrow \lambda$, $P_6 \rightarrow g \cdot$МЕСТОПОЛОЖЕНИЕ, $P_6 \rightarrow \lambda$, $P_7 \rightarrow g \cdot$РАЗМЕР, $P_7 \rightarrow \lambda$, $P_8 \rightarrow g \cdot$ИСКУССТВЕННОСТЬ, $P_8 \rightarrow \lambda$}, где $g$ – лингвистическое значение шкалы вида: {минимальный, …, малый, …, средний, …, большой, …, максимальный, λ}. Здесь λ – пустой символ.

5. Множество $G_5$ значений признака $G$ задается множеством правил порождающей грамматики: {$S \rightarrow x$, $x \rightarrow (xFx)$, $x \rightarrow xFx$, $x \rightarrow 1$ (*существующее*), $x \rightarrow 0$ (*несуществующее*), $x \rightarrow \Diamond$ (*возможное*), $x \rightarrow$ □ (*необходимое*), $F \rightarrow$ВКЛЮЧАЕТ, $F \rightarrow$ ВКЛЮЧАЕТСЯ В, $F \rightarrow$ ВКЛЮЧАЕТ И ВКЛЮЧАЕТСЯ В, $F \rightarrow$ ЧАСТИЧНО ВКЛЮЧАЕТ, $F \rightarrow$ БОЛЬШЕ ЧЕМ, $F \rightarrow$ МЕНЬШЕ ЧЕМ, $F \rightarrow$ РАВНО, $F \rightarrow$ ПОДОБНО, $F \rightarrow$ СТАНОВИТСЯ, $F \rightarrow$ ПРОИСХОДИТ ИЗ, $F \rightarrow$ ОДНОВРЕМЕННО С, $F \rightarrow$ НЕОДНОВРЕМЕННО С, $F \rightarrow$ ИМПЛИЦИРУЕТ, $F \rightarrow$ СЛЕДУЕТ ИЗ, $F \rightarrow$ СООТВЕТСТВУЕТ, $F \rightarrow$ СВЯЗАНО С}.

Все последующие уровни классификации получаются путем рекурсивного повторения предложенных пяти уровней классификации. Индекс уровня вычисляется по формуле

$$G_i = G_{\mathrm{mod}(i,5)},$$

где $i$ принадлежит множеству целых чисел.

Любому понятию или классу понятий естественного языка соответствует определенный вектор классификации [1].

Например, группе слов {иметь, получать, использовать, хранить …} соответствует вектор классификации вида [ПРЕДМЕТ \ ОТНОШЕНИЕ–СУЩЕСТВО–*X*]. Группе слов {бежать в/на/к, идти в/на/к, приближаться к, прибывать в} соответствует вектор классификации вида [МЕСТО \ ОТНОШЕНИЕ–СУЩЕСТВО–*X*]. Группе слов {видеть, смотреть на, рассматривать} – [ПРЕДМЕТ \ ОТНОШЕНИЕ–СУЩЕСТВО–*X* \\ ИДЕЯ \ ОТНОШЕНИЕ–СУЩЕСТВО–*X* \ НА НЕЖИВОМ].

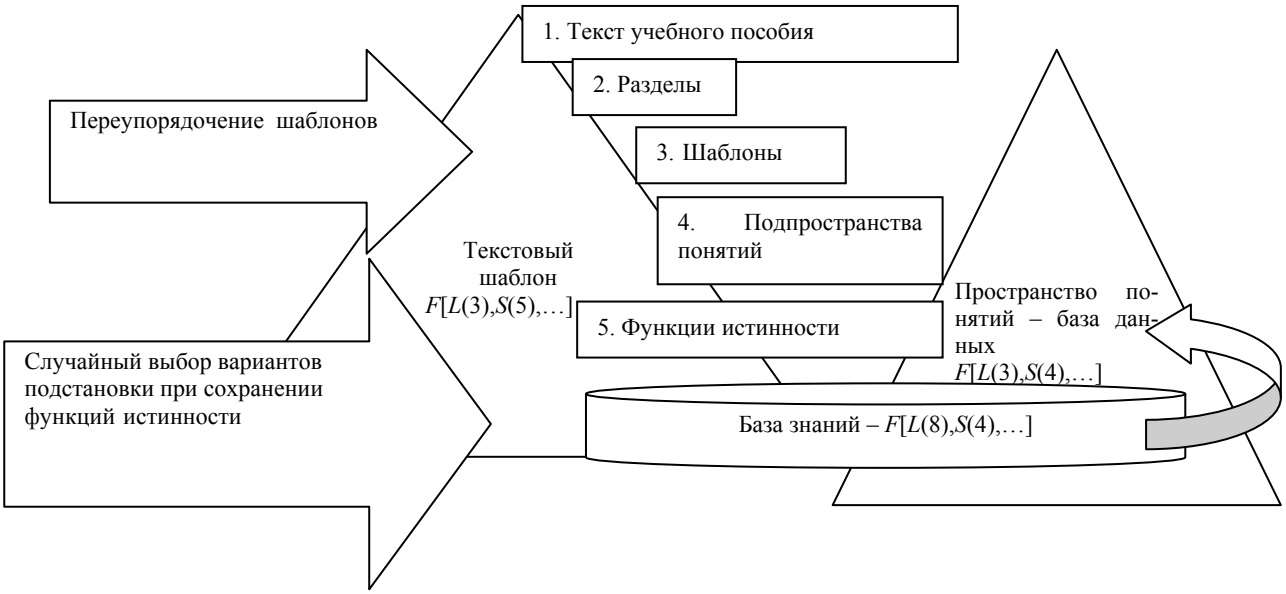


Рис. 2. Принцип автоматической генерации учебных материалов

Группе слов {мотоцикл, машина, грузовик, автобус} соответствует вектор классификации вида [ПРЕДМЕТ \ *X* \\ МЕСТО \ ОТНОШЕНИЕ–СУЩЕСТВО–*X*]. Группе слов {жадный, щедрый, экономный} соответствует вектор классификации вида [ПРЕДМЕТ \ ОТНОШЕНИЕ–*X* \\ ПРЕДМЕТ \ ОТНОШЕНИЕ–СУЩЕСТВО–СУЩЕСТВО–*X*]. Группе слов {давать, брать, покупать, продавать, дарить} соответствует вектор классификации вида [ПРЕДМЕТ \ ОТНОШЕНИЕ–СУЩЕСТВО–СУЩЕСТВО–*X*]. Так, слова «одевать», «гладить», «шить» – точки многомерного пространства, определяются координатами [ПРЕДМЕТ \ НА ПОВЕРХНОСТИ ЖИВОГО \ ДЕЙСТВИЕ] в качестве осей многомерного пространства. В свою очередь, для слов группы «аппаратное обеспечение» {монитор, клавиатура, винчестер, процессор} имеет место вектор семантических признаков вида [ПРЕДМЕТ \ У ЖИВОГО \ *X* \\ ИНФОРМАЦИЯ \ λ \ ДЕЙСТВИЕ \ СЛОЖНОЕ].

Учебный материал есть подмножество естественного языка, задаваемое шаблонами особого вида: вопрос–ответ, вопрос–варианты ответов, текст–слова к тексту и т. д. Таким образом, для создания системы автоматической генерации учебных материалов, например для уроков по английскому языку, необходимо вначале зафиксировать единицы языка, т. е. задать текстологический срез в системе естественного языка. Далее нужно описать степени свободы текстологически незафиксированных единиц языка на основе задания подпространств состояний в виде:

1) подмножеств семантической классификации слов и понятий естественного языка;

2) множеств функций истинности над подмножествами семантической классификации слов и понятий естественного языка.

Слово в предложении является распределенной системой. Так, например, слово wake … up состоит из двух элементов, распределенных в предложении.

Рассмотрим следующую модель шаблонов для автоматической генерации учебных заданий. Например, ниже приводится дерево генерации учебных заданий со ссылками на источник слов в понятийном пространстве семантической базы данных.

1. Учебное пособие *X*1 [текст].

1.1. Раздел *Y*1 [подраздел].

1.2. Раздел *Y*2 [подраздел].

1.2.1. Шаблон *Z*1 [подраздел] <текстологический текстовый шаблон>.

1.2.1.1. «Ответьте на следующие вопросы:» [позиция в предложении].

1.2.1.2. Вопрос [позиция в предложении].

1.2.1.2.1. «Можно ли найти …< в …<<?» [вариант].

1.2.1.2.2. «Находится ли …< в …<<?» [вариант].

1.2.1.2.3. «Является ли …< достопримечательностью …<<?» [вариант].

1.2.1.2.4. «Известен ли …<< таким сооружением как …<?» [вариант].

1.2.1.3. Достопримечательности [позиция в предложении] <экспорт групп слов «здания», «памятники»>.

1.2.1.3.1. Тауэр [вариант].

1.2.1.3.2. Лондонский мост [вариант].

1.2.1.3.3. Статуя Свободы [вариант].

1.2.1.3.4. Биг-Бен [вариант].

1.2.1.3.5. Царь-колокол [вариант].

1.2.1.4. Страны [позиция в предложении] <экспорт группы слов «страны»>.

1.2.1.4.1. Лондон [вариант].

1.2.1.4.1.1. Столица Великобритании [синоним].

1.2.1.4.1.2. Столица Туманного Альбиона [синоним].

1.2.1.4.2. Вашингтон [вариант].

1.2.1.4.2.1. Столица США [синоним].

1.2.1.4.3. Москва [вариант].

1.3. Раздел *Y*3 [подраздел].

В результате работы порождающей грамматики над ориентированным лесом строк [2] над шаблоном

генерации осмысленных текстов получаются тексты следующего вида:

«Англоговорящие страны имеют различные системы управления. Королева является главой Великобритании, в США глава страны – президент…»

Такие деревья генерации текстов на естественном языке можно использовать для генерации строк символов текста на выходе системы с использованием порождающих грамматик над лесом строк. Приведем примеры необходимых для этого правил порождающей грамматики, где «0» означает нулевой символ.

Как известно, стандартные порождающие грамматики над строками имеют вид четверки:

$G <S, T, N, R >$, где $S$ – начальный символ порождающей грамматики, $T$ – множество терминальных символов, $N$ – множество нетерминальных символов, $R$ – множество правил трансформации одной строки в другую.

Для порождающих грамматик над деревьями строки символов $t$ и $d$ заменяются деревьями (или ориентированным лесом – ориентированными деревьями с тождественными узлами): $t = t < t', t'',\ldots, t^n>$, где $t' = t' < t^1, t^2,\ldots, t^m>$ и т. д., $d = d < d', d'',\ldots, d^1 >$, где $d' = d' < d^1, d^2,\ldots, d^r>$ и т. д. Дерево с тождественными узлами задается в виде $d = d < d'$ $(A)$, $d''$ $(B)$, $d'''(C)$, $d''''(D),\ldots, d^1 (E)>$, где элементы ряда $A$, …, $B$, $C$, $D$, …, $E$ могут быть тождественны.

Одной из основных особенностей любой системы является наличие иерархии элементов этой системы. При этом иерархические отношения иногда могут составлять множество иерархий различных срезов рассмотрения системы. Существует понятие мультииерархических систем. Порождающие грамматики над лесом строк связаны с работой непосредственно над мультииерархическими системами данных.

Порождающая грамматика над деревьями строк строится следующим образом. Пусть $A<\ldots B<\ldots C1\rightarrow \rightarrow C2\ldots>$, …, $B'<\ldots C1'\rightarrow C2'>\ldots>$ – правило порождающей грамматики над деревьями из множества таких правил с деревьями строк терминальных символов $T$ и нетерминальных символов $N$, «→» – символ перехода одной строки в другую. $S< >$ – начальный символ порождающей грамматики над деревьями.

Углубление дерева состояний другого генерируемого дерева или леса строк состоит на каждом этапе в умножении получаемого генерируемого дерева на правило порождающей грамматики.

Можно рассмотреть также деревья, эквивалентные друг другу $A<B\{B1, B2\}, C\{C1, C2\}> = \{A<B1, C1>, A<B1, C2>, A<B2, C1>, A<B2, C2>\} = \{A<B1, C1>, A<B1, C2>, A<B2, C\{C1, C2\}>\}$, где в скобках {} отображается множество вариантов на неком уровне дерева генерации строк языка, а в скобках <> обозначается множество элементов структуры текста. Таким образом, дерево состояний системы может быть вложено в дерево элементов системы, и наоборот [4–6].

Пусть дано дерево $A <B <B'<\ldots>, B''<\ldots>, \ldots, B'''<\ldots>>, C <C'<\ldots>, C''<\ldots>, C'''<\ldots>>, \ldots, D <D'<\ldots>, D''<\ldots>, \ldots,D'''<\ldots>>>$ или коротко $A <\ldots B <\ldots B''\ldots>\ldots>$, тогда лес деревьев рассмотрим как множество деревьев с тождественными узлами на множестве узлов этих деревьев: $F<A<\ldots B<\ldots B''(=L1)\ldots>\ldots>, X<\ldots Y<\ldots Y''(=L1)\ldots>\ldots>,\ldots>$, где $L1$ – тождественный узел первых двух деревьев вышеприведенного примера.

Принцип свертки или сложения образов заключается в следующем: семантически схожие элементы – узлы деревьев – объявляются тождественными, в случае наличия нескольких вариантов свертки строится дополнительное подпространство возможных состояний системы – результата сложения деревьев элементов системы и порождения деревьев состояний системы.

Алгоритмический шаг по выбору одного из синонимов в шаблон:

1. $A$ […] → $C$.
1.1. … → 0.
1.2. $C$ [синоним] → 0.
1.3. … → 0.

Алгоритмический шаг по выбору одного из вариантов подстановки в шаблон:

1. $A$ […] → $C$.
1.1. … → 0.
1.2. $C$ [вариант] → 0.
1.3. … → 0.

Алгоритмический шаг по представлению дерева строк в последовательность строк:

1. $A$ […] → 0.
1.1. $B$ [раздел] → 0.
1.2. $D$ [раздел] → 0.
1.3. $C$ [раздел] → 0.
2. 0 → $B$.
3. 0 → $D$.
4. 0 → $C$.

В результате обработки данного дерева символов на основе расширенных порождающих грамматик над лесом строк на основе приводимых ниже правил получаются следующие строки символов:

«Является ли Тауэр достопримечательностью Лондона?»

«Находится ли статуя Свободы в Москве?»

Важным аспектом является генерация учебных текстов. Для генерации простейших учебных текстов будем использовать дерево генерации текста:

1. Учебное пособие $X2$.
1.1. Раздел $Y1$ [раздел].
1.2. Раздел $Y2$ [раздел].
1.2.1. Шаблон $Z1$ [раздел].
1.2.1.1. Лицо [позиция в предложении].
1.2.1.1.1. Я [вариант].
1.2.1.1.2. Мой друг [вариант].
1.2.1.1.3. Моя подруга [вариант];
1.2.1.1.4. Мой одногруппник [вариант].
1.2.1.1.5. Мой дядя [вариант].
1.2.1.2. Аспект [позиция в предложении].
1.2.1.2.1. Имя [позиция в предложении].
1.2.1.2.1.1. Фамилия.
1.2.1.2.1.1.1. Связка [варианты].
1.2.1.2.1.1.1.1. Имя … –.
1.2.1.2.1.1.1.2. … зовут.
1.2.1.2.1.1.1.3. … –.

1.2.1.2.1.1.1.4. Фамилия … –.
1.2.1.2.1.1.2. Типы книг.
1.2.1.2.1.1.2.1. Иванов.
1.2.1.2.1.1.2.2. Петров.
1.2.1.2.1.1.2.3. Сидоров.
1.2.1.2.1.1.2.4. Браун.
1.2.1.2.2. Год рождения [позиция в предложении].
1.2.1.2.3. Знак зодиака [позиция в предложении].
1.2.1.2.4. Характер [позиция в предложении].
1.2.1.2.5. Отношения [позиция в предложении].
1.2.1.2.6. Занятия [позиция в предложении].
1.2.1.2.6.1. Книги.
1.2.1.2.6.1.1. Позитивная модальность [варианты].
1.2.1.2.6.1.1.1. Обожать.
1.2.1.2.6.1.1.2. Любить.
1.2.1.2.6.1.1.3. Нравится.
1.2.1.2.6.1.1.4. Часто.
1.2.1.2.6.1.1.5. Постоянно.
1.2.1.2.6.1.2. Действия с книгами.
1.2.1.2.6.1.2.1. Читать/Читает.
1.2.1.2.6.1.2.2. Перечитывать/перечитывает.
1.2.1.2.6.1.2.3. Просматривать/просматривает.
1.2.1.2.6.1.3. Типы книг.
1.2.1.2.6.1.3.1. Книги.
1.2.1.2.6.1.3.2. Классику.
1.2.1.2.6.1.3.3. Детектив.
1.2.1.2.6.1.3.4. Сказки.
1.2.1.2.6.2. Музыка.
1.2.1.2.6.3. Кино.

В результате генерации осмысленного подмножества естественного языка получаются тексты следующего вида:

«Меня зовут Иван. Моя фамилия – Иванов. Я родился первого октября 1980 г. Мне 30 лет. Мой знак зодиака – Весы. Я люблю классическую музыку…»

При этом строки на узлах деревьев разнородных данных должны быть объединены в лес данных, частично как подмножества понятийного пространства естественного языка, частично как подмножества иерархии шаблонов электронного учебного курса.

На основе порождающих грамматик над лесом строк возможна дополнительная генерация дерева текстовых шаблонов с добавлением семантического шума и порождение текста на естественном языке, включающем информацию в иерархической системе текста, сгенерированную на основе знаний базы данных. Так, например, вместо фразы «я люблю читать детективы» может быть употреблена разговорная фраза не приведенного вида: «я не прочь четануть детективчик» с той же формально-базовой семантикой.

В заключение необходимо отметить, что классификация понятий естественного языка может служить источником лексических единиц для составления шаблонов генерации осмысленных текстов, которые можно усложнять посредством добавления семантического шума на основе расширенных порождающих грамматик над лесом строк и деревьев разнородных данных.

**Библиографические ссылки**

1. Личаргин Д. В. Методы и средства порождения семантических конструкций естественно языкового интерфейса программных систем : автореф. дис. … канд. техн. наук. Красноярск, 2004.

2. Личаргин Д. В. Порождение дерева состояний на основе порождающих грамматик над деревьями строк // Вестник СибГАУ. 2009. Вып. 4 (25). С. 33–37.

3. Личаргин Д. В. Операции над семами слов естественного языка в машинном переводе // Тр. конф. молодых ученых / Ин-т вычисл. моделирования СО РАН. Красноярск, 2003. С. 23–31.

4. Агамджанова В. И. Контекстуальная избыточность лексического значения слова. М. : Высш. шк., 1977.

5. Апресян Ю. Д. Идеи и методы современной структурной лингвистики. М. : Наука, 1966.

6. Вердиева З. Н. Семантические поля в современном английском языке. М. : Высш. шк., 1986.

K. V. Safonov, D. V. Lichargin

## SOME PRINCIPLES OF EDUCATIONAL MATERIALS AUTOMATIC GENERATION BASED ON DATABASES AND A LINGUISTIC CLASSIFICATION

*In the work models and means of meaningful subset of the natural language generation for educational courses are considered. In particular, a semantic notional space of the words of the language is assigned. The purpose of text generation model creation for the educational courses in English language, is set, the tasks of its application, based on generative grammars over oriented forest of strings, is formulated, and the conclusion about the specific features and the structure of the educational courses generation model, is made.*

*Keywords: natural language generation, semantic features, language words and notions classification, educational materials generation.*