УДК 004.9

Вестник СибГАУ
2014. № 5(57). С. 93–100

РАЗРАБОТКА СИСТЕМЫ РЕКУРСИВНЫХ ПОРОЖДАЮЩИХ ГРАММАТИК ДЛЯ РЕШЕНИЯ ЗАДАЧИ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ИНТОНАЦИОННЫХ ШАБЛОНОВ ЯЗЫКОВЫХ ВЫРАЖЕНИЙ

Д. В. Личаргин[1], А. И. Трушакова[1], К. В. Сафонов[2], Е. П. Бачурина[1]

[1]Сибирский федеральный университет
Российская Федерация, 660074, г. Красноярск, ул. академика Киренского, 28
[2]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: lichdv@hotmail.ru

Рассматривается проблема, заключающаяся в разработке нового класса порождающих грамматик, обеспечивающих лучшее качество генерации строк и внутреннюю упорядоченность правил. Цель работы состоит в теоретическом обосновании и построении модуля разработанной ранее программы по автоматической генерации транскрипции на различных европейских языках, обеспечивающего добавление интонационной разметки, что даст возможность обучающемуся более качественно изучать язык, его фонетический строй и особенности. Кроме того, важным является решение проблемы определения модальности высказывания – различение повествования, вопроса и восклицания, что должно обеспечить возможность говорить и выражать множество эмоций на иностранном языке без предварительного знания слов и правил. Данный тип порождающих грамматик – рекурсивные порождающие грамматики – должен позволить решать более широкий класс задач благодаря использованию разбиения всего класса правил порождающих грамматик на уровни и разделы, что будет отражено в структуре соответствующих правил. Методы реализации поставленной цели состоят в гибридизации порождающих грамматик и подстрок команд в форме особых нетерминальных символов языка порождающих грамматик. В результате исследования сформулирован принцип организации множества деревьев, состояний порождаемой строки высказывания на естественном языке. Разработанный подход позволяет обеспечить выполнение алгоритмов генерации строк естественного языка с учетом использования командных нетерминальных символов порождающих грамматик. Заявленный подход к расширению синтаксиса правил порождающих грамматик может позволить обеспечить более сложную и качественную трансформацию генерируемой строки алфавита символов с учетом структурирования деревьев состояния строки на уровне рекурсивной генерации. Такой инструментарий компьютерной лингвистики может найти свое применение в качестве средства моделирования и анализа естественных языков, в частности, в целях генерации осмысленной речи и осуществления языковых трансформаций. Это позволит решать более сложные классы проблем в компьютерной лингвистике и разработке лингвистического программного обеспечения.

Ключевые слова: порождающие грамматики, рекурсивные порождающие грамматики, компьютерная лингвистика, искусственный интеллект, генерация транскрипций.

Vestnik SibGAU
2014, No. 5(57), P. 93–100

DEVELOPMENT OF RECURSIVE GENERATIVE GRAMMARS SYSTEM FOR SOLVING THE PROBLEM OF AUTOMATIC GENERATION OF LANGUAGE PHRASES INTONATION PATTERNS

D. V. Lichargin[1], A. I. Trushakova[1], K. V. Safonov[2], E. P. Bachurina[1]

[1]Siberian Federal University
28, Kirenskiy str., Krasnoyarsk, 660074, Russian Federation
[2]Siberian State Airspace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: lichdv@hotmail.ru

This paper considers the problem of developing a new class of generative grammars, providing higher quality of string generation and internal order of the rules. The purpose is to justify theoretically and to develop the module in the

previously developed system for automatic generation of transcription in various European languages, which allows adding the intonation patterns, which will enable the student to study the language, its phonetic structure and features in a more effective way. In addition, it is important to solve the problem of modality identification, the distinction between narrative sentences, questions and exclamations that will provide an opportunity to speak and express many types of emotions in a foreign language without preliminary knowledge of words and rules. This type of generative grammars – "recursive generative grammars" – should allow solving a wider class of problems through the use of dividing the entire class of rules within generative grammars to levels and sections, what will be demonstrated in the structure of the relevant rules. Methods to accomplish the above purpose consist in hybridization of generative grammar and substring of command in the form of nonterminals of generative grammar. As a result of work a principle organizing a set of trees of the states of generated string expressions in the natural language has been developed. The developed approach allows the implementation of the algorithms for generating the strings of the natural language from concerning the application of nonterminal characters of generative grammars. The proposed approach to extending the syntax of the generative grammar rules can allow providing a more complex and qualitative transformation of generated string of the characters alphabet considering the structuring process for the trees of states on the levels of recursive generation. These tools of computational linguistics can be applied as the means of modeling and analyzing the natural languages, in particular, in order to generate meaningful speech and carry out language transformations. This will allow solving more complex problems in computational linguistics and the development of linguistic software.

Keywords: generative grammars, recursive generative grammars, computational linguistics, artificial intelligence, phonetic transcription generation.

Введение. В настоящее время порождение речи компьютером является крайне важным. Значительная часть программных приложений, связанных с генерацией текстов на естественном языке, работает с информацией на естественном языке.

В работе рассматривается проблема, заключающаяся в разработке нового класса порождающих грамматик, обеспечивающих лучшее качество генерации строк и внутреннюю упорядоченность правил.

Проблематика, связанная с генерацией порождающими грамматиками, давно и широко исследуется различными авторами, в частности Хомским, Монтегю и др. [1; 2]. Исследования проводятся на стыке таких наук, как информатика, лингвистика, компьютерная лингвистика, математика и математическая логика [3–6].

Однако улучшение качества генерации строк требует дополнительных исследований в рамках новых классов порождающих грамматик.

Основная идея работы состоит в построении принципа работы рекурсивных порождающих грамматик на основе введения потенциально генерируемых команд-примечаний и в применении этой модели к проблеме генерации фонетических шаблонов в образовательных целях.

Целью данной работы является разработка модели системы рекурсивных порождающих грамматик, которая позволит обучающимся произносить предложения на незнакомом или малознакомом языке с правильной интонационной окраской.

Задачи данной работы заключаются в следующем:

1. Разработка специализированного синтаксиса, позволяющего реализовать базовый функционал данного типа порождающих грамматик, при этом не противоречащий общему синтаксису предшествующих типов порождающих грамматик.

2. Разработка принципа построения правил порождающих грамматик для построения интонационной разметки в отдельном слове, словосочетании, выражении.

3. Разработка алгоритмической реализации данного класса порождающих грамматик.

4. Разработка набора правил для данной модели на основе предложенного функционала рекурсивных порождающих грамматик (РПГ), что позволит сгенерировать интонационные обозначения целевых выражений для конечного пользователя.

Новизна работы состоит в использовании общей схемы работы порождающих грамматик с привлечением синтаксиса командных подстрок.

Рассматриваемая в настоящей работе модель должна позволить обучающимся повысить качество чтения текста на иностранном языке с учетом маркеров изменения относительной высоты тона в слоге, слове и целом высказывании. Это позволит определять законченность или незаконченность высказывания, завершенность мысли и эмоциональный оттенок речи.

Кроме того, система генерации транскрипции и интонационной разметки во многом решает проблему описания модальности высказывания – различение повествования, вопросы и восклицания. Указанные методы должны обеспечить возможность говорить и выражать множество эмоций на иностранном языке без предварительного знания слов и правил.

Принципы построения рекурсивных порождающих грамматик. Порождающая грамматика позволяет выводить (порождать) цепочки языка из некоторой начальной цепочки с помощью определенных правил замены (или правил подстановки).

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 и R – множество правил трансформации одной строки в другую.

Порождение есть пошаговый процесс, в котором на каждом шаге из цепочки, уже полученной на предыдущем шаге (в частности, из начальной), можно путем применения к ней правил замены получить новую цепочку [7; 8].

Разработка системы поддержки рекурсивных порождающих грамматик. На основе порождающих грамматик были разработаны правила для генерации транскрипции иностранных языков, а также функции, реализующие генерацию верной и неверной транскрипций, что решает проблему составления вариантов ответов к тестам электронных курсов.

Создание таких правил решало актуальную проблему, появившуюся в связи с широким распространением подходов по проверке знаний учеников в электронной форме.

Была создана программа генерации правильной и неправильной транскрипции и создания на их основе заданий с верными и неверными вариантами ответов для учебных тестов. Предложенная программа позволяет выбирать язык генерации из предложенных, перевести введенное слово или текст в корректную и некорректную транскрипцию, с помощью которых возможно составить тесты, что может значительно упростить процесс контроля успеваемости у студентов и школьников, изучающих иностранный язык, в частности, английский. При порождении фраз порождающей грамматикой используются различные синтаксические формы, в частности, указание на параллельную замену подстрок (рис. 1).

В перспективе на основе использования рекурсивных порождающих грамматик с привлечением семантических моделей языка можно более эффективно решать задачи порождения грамматически и семантически осмысленной речи. С другой стороны, на основе расширения традиционных порождающих грамматик можно повышать эффективность различных аспектов для более узких классов задач от генерации фонетической транскрипции до работы трансляторов языков программирования высокого уровня.

Также необходимо добавить, что в основе предложенной модели лежит программа «Генератор классификаций», которая представляет собой визуализацию процесса работы порождающих грамматик в форме дерева вывода для некоторой части области знаний на уровне интерфейса пользователя. Такая схема вывода строится на базе правил, описывающей соответствующие релевантные классы объектов или их состояний и переходы между объектами / состояниями, известными в этой области, заданными наборами правил порождающих грамматик (рис. 2).

Отметим, что предшествующие типы порождающих грамматик не всегда справляются с задачей учета более широкого контекста и нуждаются в увеличении качества генерируемых строк, а также в устранении недостаточного упорядочения внутренней структуры множеств правил. Необходимо учесть традиционное отсутствие их деления на классы или группы. Предлагается создание данного типа порождающих грамматик – рекурсивные порождающие грамматики, который позволит решать данный класс задач. Общеизвестно, что рекурсия заключается в принципе определения, описания или вычисления функции или набора правил, содержащего аналогичные структуры внутри самой этой функции, т. е. когда объект вызывает сам себя, точнее, аналогичный объект на другом этапе вычисления.

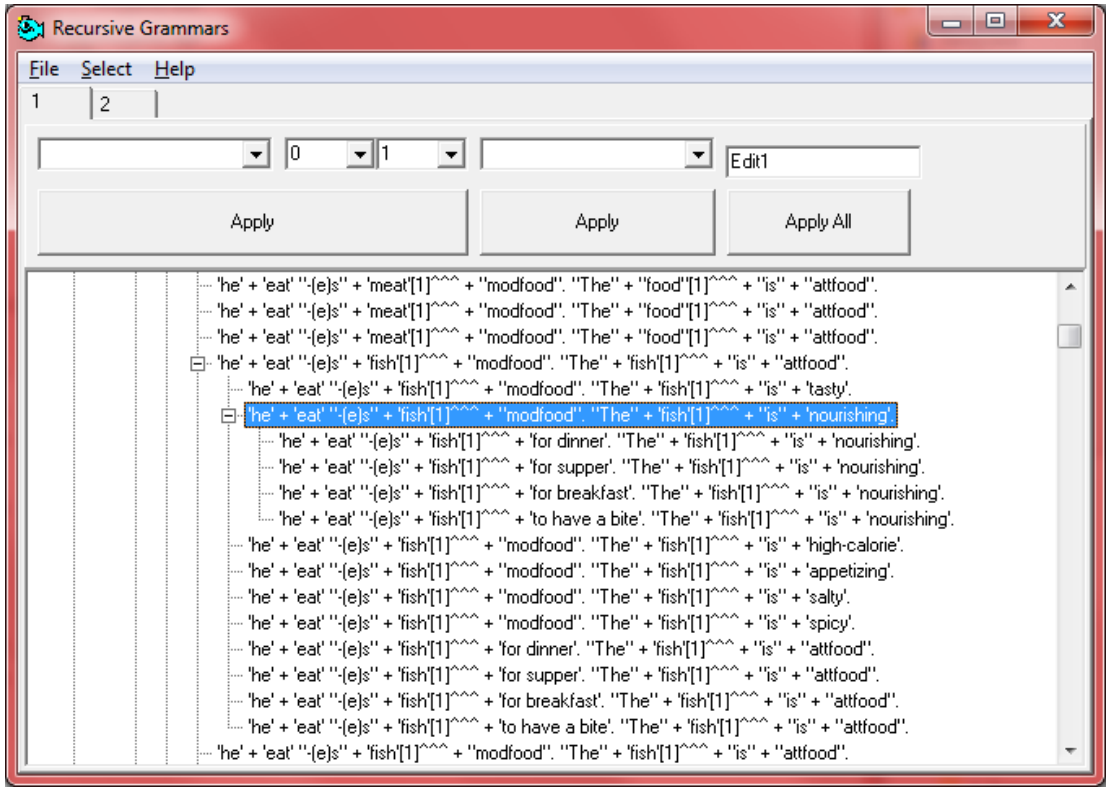

Рис. 1. Пример генерации осмысленных фраз языка с параллельной заменой соответствующих друг другу подстрок в программе «Генератор транскрипций»

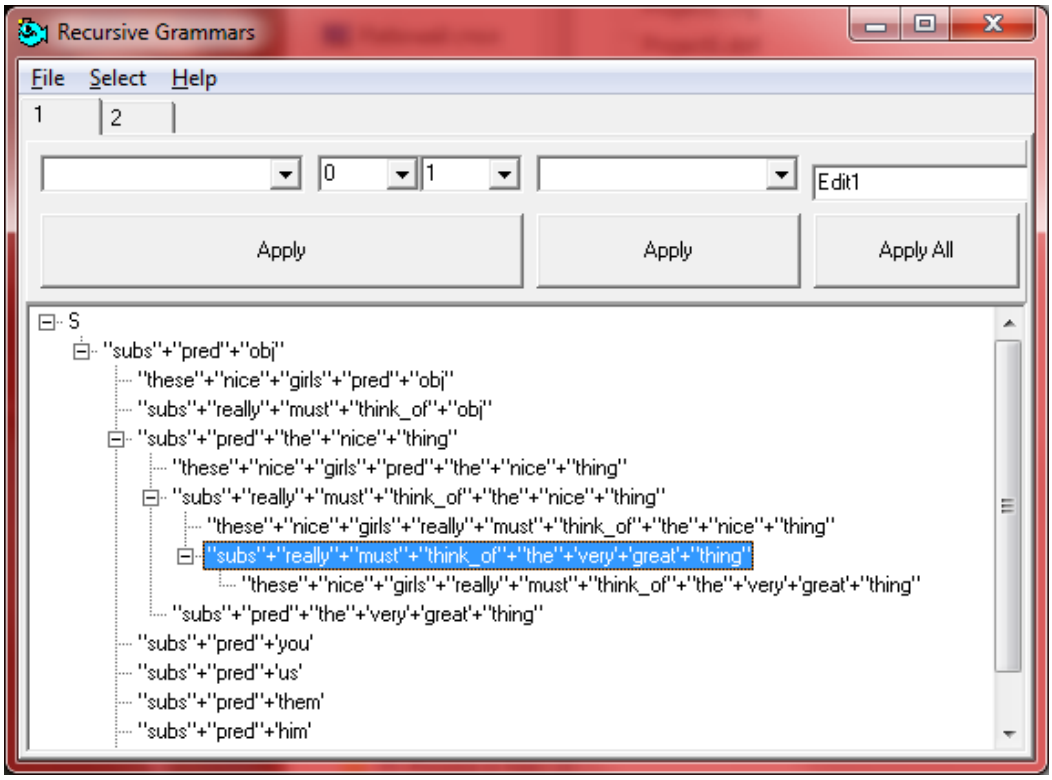

Рис. 2. Пример действия программы «Генератор классификаций», модуль Recursive Grammars

При этом сами правила порождающих грамматик, имеющих вид «подстрока$_1$» + «подстрока$_2$» + … + «подстрока$_n$» → «подстрока$_1$'» + «подстрока$_2$'» + … + «подстрока$_n$'», будут иметь следующие добавления в синтаксисе внутреннего описания правил, а именно, после любой строки может находиться команда вида: «строка$_i$» + «строка$_{i+1}$» + … {название_базы_правил, глубина_генерации, количество_строк}.

Приведем пример использования предлагаемого вида правил рекурсивной порождающей грамматики:

«The» + «nice» + «girl» + «like» + + «-(e)s»{endings,4,2} + «to» + «dance» + «обстоятельство места».

Результатом применения вложенного набора правил порождающей грамматики на втором уровне рекурсии будет строка вида

«The» + «nice» + «girl» + «likes» + «to» + «dance» + + «обстоятельство места».

Приведем второй пример синтаксиса правил порождающей грамматики рассматриваемого класса:

«The» + _ «nice» + / «girl» + _ «модальность»{modality,2,1} + / «dance» + «in» + «the» + «атрибут здания/организации» + \ «club».

Результатом применения вложенного набора правил порождающей грамматики на втором уровне рекурсии будет строка вида

«The» + _ «nice» + / «girl» + _ «starts» + >«-ing» + + / «dance» + «in» + «the» + «атрибут здания/организации» + \ «club».

И далее получаем строку вида:

«The» + _ «nice» + / «girl» + _ «starts» + / «dance» «-ing»{endings,4,2} + «in» + «the» + «атрибут здания/организации» + \ «club».

Далее при применении правила базы правил endings получаем строку

«The» + _ «nice» + / «girl» + _ «starts» + / «dancing» + «in» + «the» + «атрибут здания/организации» + \ «club» и далее получаем строку вида

«The» + _ «nice» + / «girl» + _ «starts» + / «dancing» + + «in» + «the» + «youth» + \ «club» или

«The» + _ «nice» + / «girl» + _ «starts» + / «dancing» + + «in» + «the» + «sport» + \ «club» и т. п.

Фрагменты схемы алгоритмической реализации решения проблемы генерации фонетической транскрипции и разметки приводится ниже.

На сегодняшнем этапе реализован первый этап построения генератора материалов (в частности, по методу И. Франка) а именно, программа генератор транскрипции [9]. Данная программа использует следующие базы данных:

1. Порождающие грамматики следующего вида:

– русская транскрипция:

и 'i
жи zhi
же zhe
жа zha
жо zho
…

– английская транскрипция:

ii ии
ie И:
io иОу
iu иэ
ei Еи
ee и:
ea и:
…

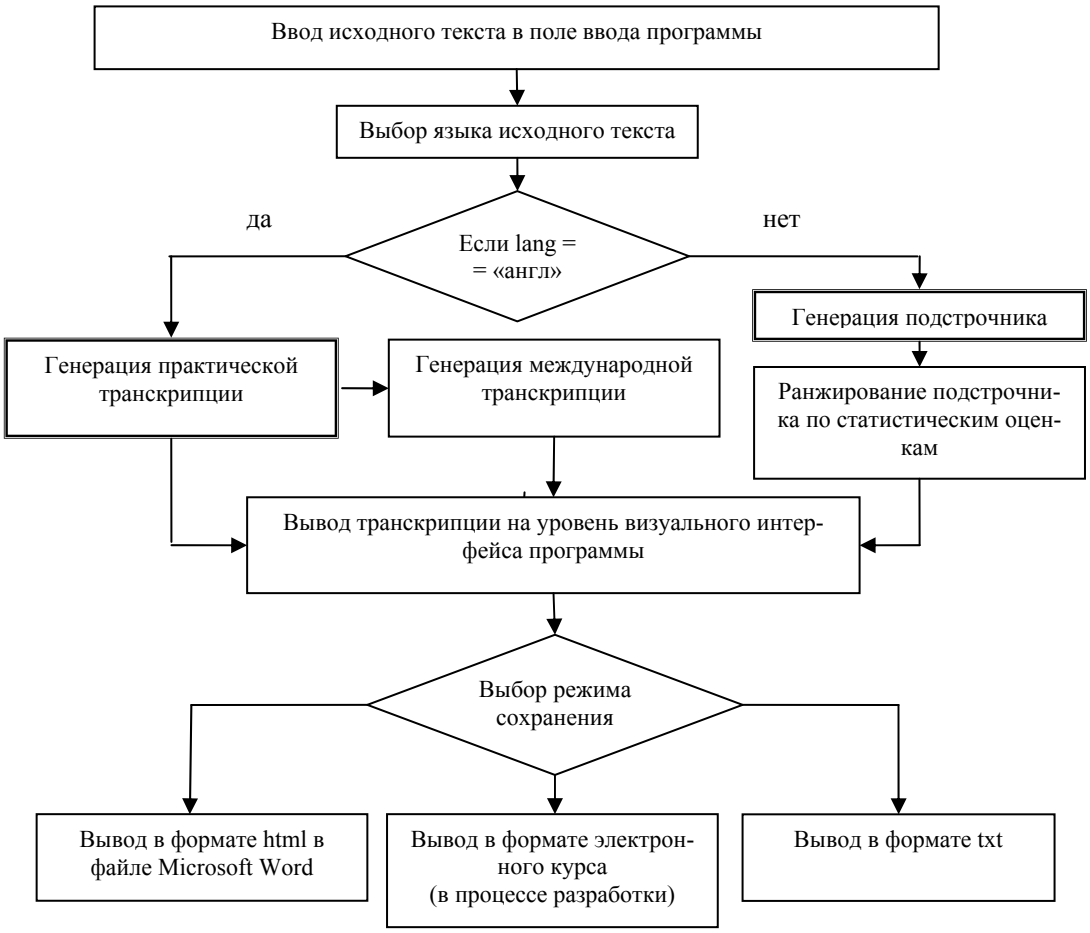

Рис. 3. Алгоритм работы программы «Генератор транскрипции»

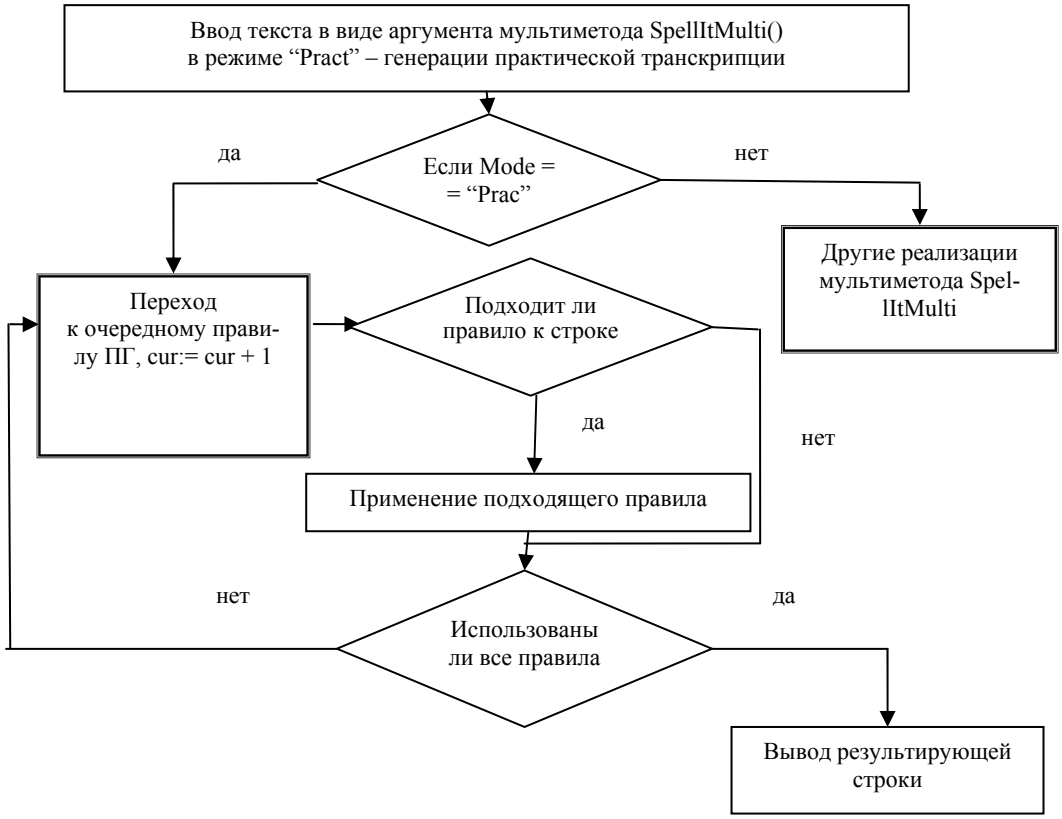

Рис. 4. Алгоритм генерации практической транскрипции на основе правил

2. База неправильной транскрипции для составления тестов:

ory opy
ury yy
ir ир
er ep
ar ap
…

3. База данных транскрипции слов (на 10 тыс. слов):

state – стЕйт
significant – сэгнИфикэнт
Siberia – сайбИ:риэ
Siberian – сайбИ:риэн
Federal – фЕдэрэл
research – рисЭ:ч
neural – нйУ:рэл
network – нЕтВэ:к
net – нЕт
…

В правилах порождающей грамматики слева стоит символ, который должен быть преобразован в символ, стоящий справа.

Различные типы функций, реализующие различные принципы обработки строк, реализуются на основе обобщающих методов (рис. 3, 4), что упрощает работу по трансформации строк.

На рис. 5 представлены рекурсивные уровни генерации строк естественного языка в форме деревьев состояний строки.

В рамках данного исследования система генерации «леса» состояний генерируемой строки описывается моделью последовательного вызова правил определенного типа для конкретных уровней генерации с возможностью семантизации этих уровней. Данный подход является частью работы по формализации естественного языка с акцентом на его семантику и решение проблемы генерации осмысленной речи [10–15].

Стандартный подход к решению проблемы представления правил порождающих грамматик обычно сводится к использованию неупорядоченного множества записей базы правил, что дает возможность корректировать результаты генерации интуитивно, а не на основе систематизации представления более или менее качественного набора правил. Корреляция между степенью упорядоченности правил порождающей грамматики и процентным соотношением правильно генерируемых, транскрипционных и фонетических правил обусловлено фактором сложности и перепутанности логических соотношений в таких правилах. Таким образом, база правил лингвистического программного обеспечения может быть представлена в форме отдельных подмножеств, отладка которых по отдельности сокращает количество ошибок генерации системы в целом. Множества правил структурируются по смысловым категориям, например, фонетические, морфологические, грамматические, лексические и т. п.

Необходимо отметить, что особенности, преимущества и потенциальные недостатки в работе системы рекурсивных порождающих грамматик должны быть зафиксированы, обработаны и интерпретированы на основе многочисленных экспериментов. Распределение правил по классам в зависимости от их функционального назначения должно обеспечить требуемые результаты для решения классов более сложных задач с учетом широкого структурированного контекста. Важным этапом работы по исследованию новых классов порождающих грамматик на основе упорядочения их множества и применения «командных» нетерминальных символов является решение конкретных лингвистических задач в качестве показательных примеров, дающих возможность осуществить отладку ошибок на множестве правил рекурсивных порождающих грамматик, устранить избыточность правил, использовать высокий уровень обобщения при их построении.

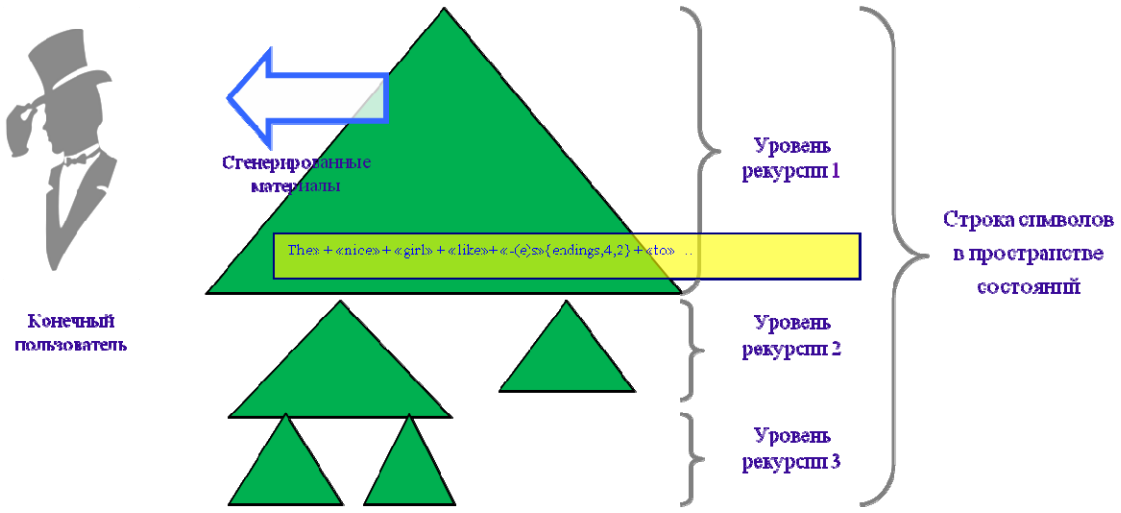

Рис. 5. Общая схема работы рекурсивных порождающих грамматик

Заключение. Таким образом, предложена модель системы рекурсивных порождающих грамматик для учебных целей. Предложены принципы синтаксического оформления, позволяющие реализовать базовый функционал данного типа порождающих грамматик, предложен алгоритм применения правил порождающих грамматик для построения интонационной разметки в отдельном слове, словосочетании, выражении, проведено обоснование алгоритмической реализации данного класса порождающих грамматик. Приведены примеры правил для системы генерации строк естественного языка с рекурсией, вызываемой командными нетерминальными символами порождающей грамматики.

Генерация фонетической разметки является примером предметной области, в которой сложно получить требуемый результат с использованием только традиционных порождающих грамматик, включая контекстно-свободные порождающие грамматики, что подчеркивает актуальность заявленной темы.

При создании программы на основе предложенной модели необходимо решить следующие проблемы: увеличение порождающей мощности, устранение недостатка внутренней структуры множеств правил, использование деления на классы или группы. Программа позволит генерировать интонационную разметку, что значительно упростит процесс изучения учащимися новых языков.

Разрабатываемая программная система позволит обрабатывать большие объемы текста для их использования в качестве материалов к уроку по иностранному языку, самостоятельной работы и др. Данная система позволит решить задачи широкого круга пользователей, например, студентам – ликвидировать пробелы в знаниях по иностранному языку, в особенности в фонетике иностранного языка; изучить новый иностранный язык и говорить с правильной интонацией уже на первом занятии; преодолеть языковой барьер в изучении основ разговорного иностранного языка в высших учебных заведениях; позволит туристам, посещающим европейские страны, понять культуру и традиции другой страны, практиковать иностранный язык без долгого предварительного изучения правил, грамматики, лексики, фонетики, орфографии.

Благодарности. Работа частично выполнена при поддержке Министерства образования и науки Российской Федерации, проект Б 112/14.

Acknowledgment. This work was financially supported by the Ministry of Education of the Russian Federation, the project Б 112/14.

Библиографические ссылки

1. Carnap R. Meaning and Necessity // A Study in Semantics and Modal Logic.1956.

2. Chomsky N. Syntatic Structures: Mouton de Gruyter. 2002.

3. Хомский Н. Синтаксические структуры // Новое в зарубежной лингвистике. Вып. V. М., 1962.

4. Automatic Expansion of Domain-Specific Lexicon by Term Categorization / H. Avancini [et al.] // ACM Translation on Speech and Language Processing. 2006. Vol. 3, No. 3. P. 1–30.

5. Evaluating Discourse Understanding in Spoken Dialogue Systems / R. Higashinaka [et al.] // ACM Translation on Speech and Language Processing. 2004. Vol. 1. P. 1–20.

6. Towards Efficient Human Machine Speech Communication: The Speech Graffiti Project / S. Tomko [et al.] // ACM Translation on Speech and Language Processing. 2005. Vol. 2. No. 1.

7. Сафонов К. В. О возможности вычислительного распознавания контекстно-свободных грамматик // Вычислительные технологии. 2005. Т. 10, № 4. С. 91–98.

8. Сафонов К. В., Егорушкин О. И. О синтаксическом анализе и проблеме В. М. Глушкова распознавания контекстно-свободных языков Хомского // Вестник Томского государственного университета. 2006. № 17. С. 63.

9. Свидетельство о государственной регистрации базы данных № 2014615928. Генератор транскрипции / Е. П. Бачурина, Д. В. Личаргин. Заявл. 14.04.2014 ; опубл. 05.06.2014.

10. Личаргин Д. В. Порождение дерева состояний на основе порождающих грамматик над деревьями строк // Вестник СибГАУ. 2010. № 1(27). С. 57–59.

11. Личаргин Д. В., Бачурина Е. П. Обобщенная иерархическая структура учебного электронного курса и рассмотрение на ее основе электронных курсов обучения английскому языку // Информатизация образования и науки. 2012. № 3(15). С. 20–36.

12. Статистические методы анализа естественного языка как способ повышения эффективности его генерации на основе семантических шаблонов / Д. В. Личаргин [и др.] // Информатизация образования и науки. 2014. № 4(24). С. 92–103.

13. Бачурина Е. П., Трушакова А. И., Личаргин Д. В. Информационная система генерации тестовых заданий по фонетике иностранного языка для студентов технических вузов // Современные инновации в науке и технике : сб. науч. тр. 4-й Междунар. науч.-практ. конф. Т. 1 / Юго-Зап. гос. ун-т. Курск, 2014. С. 115–120.

14. Бачурина Е. П. Разработка программы генерации учебных материалов по иностранному языку на основе порождающих грамматик Хомского // Современные инструментальные системы, информационные технологии и инновации : сб. науч. тр. XI-й Международной науч.-практ. конф. Т. 1 / Юго-Зап. гос. ун-т. Курск, 2014. С. 204–207.

15. К вопросу об упорядочении многоуровневой семантической сети на дереве семантической классификации / Д. В. Личаргин [и др.] // Вестник СибГАУ. 2014. № 2. С. 44–50.

References

1. Carnap R. Meaning and Necessity. A Study in Semantics and Modal Logic. 1956.

2. Chomsky N. Syntatic Structures. Berlin: Mouton de Gruyter, 2002.

3. Chomsky N. [Syntax structures] *Novoe v zarubezhnoy lingvistike*, 1962, vol. V (In Russ.).

4. H. Avancini, A. Lavelli, F. Sebastiani, R. Zanoli. Automatic Expansion of Domain-Specific Lexicon by Term Categorization. *ACM Translation on Speech and Language Processing*. 2006, vol. 3, no. 1, p. 1–30.

5. Higashinaka R., Miyazaki N., Nakano M., Aikawa K. Evaluating Discourse Understanding in Spoken Dialogue Systems. *ACM Translation on Speech and Language Processing*. 2004, vol. 1, p. 1–20.

6. Tomko S., Harris T. K., Toth A., Sanders J., Rudnicky A., Rosenfeld R. Towards Efficient Human Machine Speech Communication: The Speech Graffiti Project. *ACM Translation on Speech and Language Processing.* 2005, vol. 2, no. 1.

7. Safonov K. V. [The possibility of computational recognition of context-free grammars] *Vichislitelnie tekhnologii*. 2005, vol. 10, no. 4, p. 91–98 (In Russ.).

8. Safonov K. V., Egorushkin O. I. [Syntactical analysis and problem Glushkov recognition of Chomsky's context-free languages] *Bulletin of Tomsk State University*. 2006, no. 17, p. 63.

9. Bachurina E. P., Lichargin D. V. *Generator transkripcii* [Generator of transcription]. Certificate of state registration of software No. 2014615928 of June 5, 2014 (In Russ.).

10. Lichargin D. V. [Generation of tree of states based on generative grammars over the trees of string]. *Vestnik SibGAU*. 2010, no. 1 (27), p. 57–59 (In Russ.).

11. Lichargin D. V., Bachurina E. P. [Generalized hierarchical structure of the educational e-learning course and base principles of e-learning in teaching English]. *Informatizacia Obrazovaniya I Nauki.* 2012, no. 3 (15), p. 20–36 (In Russ.).

12. Lichargin D. V., Maglinets A. Y., Ribkov M. V., Bachurina E. P. [Statistical methods for the analysis of natural language as a way to improve the efficiency of its generation based on semantic templates]. *Informatizacia Obrazovaniya I Nauki*. 2014, no. 4 (24), p. 92–103 (In Russ.).

13. Bachurina E. P., Trushakova A. I., Lichargin D. V. [Information system for generating test tasks on phonetics of foreign language for students of engineering specialty]. *Sovremennie innovatsii v nauke i tehnike: sbornik nauchnikh trudov 4 mezhdunarodnoy konferentsii* [Modern innovations in science and technology: Proceedings of the 4th International Conference]. 2014, vol. 1, p. 115–120 (In Russ.).

14. Bachurina E. P. [Development of a system for generation of training content in a foreign language based on the Chomsky's generative grammar]. *Sovremennie instrumentalnie sistemi, informatsionnie tekhnologii i innovatsii*: *sbornik nauchnikh trudov XI mezhdunarodnoy konferentsii* [Modern instrumentation systems, information technology and innovation: Proceedings of the XI-th International Conference]. Kursk, 2014, vol. 1, p. 204–207 (In Russ.).

15. Lichargin D. V., Safonov K. V., Egorushkin O. I., Bachurina E. P. [About the issue of ordering a multilevel semantic web on the tree of semantic classification]. *Vestnik SibGAU*. 2014, no. 2, p. 44–50 (In Russ.).