Синтез визуальных моделей текста

Кучуганов В.Н. (kuchuganov@istu.ru)

ГОУ ВПО "Ижевский государственный технический университет"

Визуальные модели текста служат для проверки правильности морфологического, синтаксического и семантического анализа текста на естественном языке. В работе предлагается подход к решению задачи визуального моделирования, заключающийся в том, что модель знаний о грамматике языка изначально ориентирована на графическое отображение языковых конструкций и описываемых сюжетов.

Описанные модели заложены в учебную систему KG (Knowledge's Guidebook) для разработки онтологий и экспертных систем, используемую при обучении студентов.

Введение

Представление текстовой информации в виде схем, графиков, диаграмм, статических и динамических геометрических моделей – это форма, которая позволяет добиться однозначности истолкования грамматических конструкций и проверить адекватность внутренних машинных моделей текста тому смыслу, который закладывал автор, хотя следует признать, что не всякий текст можно изобразить графически.

Задача синтеза граф-схемы заданного текста и модификации текста при корректировке пользователем этой схемы является актуальной, в частности, при изучении грамматики естественного языка (ЕЯ), при исследовании алгоритмов семантического анализа и контекстного перевода, в задачах управления контекстами при работе с удаленными приложениями [Попов, 1987; Гаврилова, 2000].

Ниже предлагается подход к решению задачи визуального моделирования текстов на ЕЯ, заключающийся в том, что модель знаний о предметной области изначально ориентирована на графическое отображение языковых конструкций и описываемых сюжетов. В качестве инструмента разработки онтологий предлагается система KG (Knowledge Guide-book), разработанная на кафедре «Автоматизированные системы обработки информации и управления» Ижевского государственного технического университета [Кучуганов и др., 2001]. Благодаря большей степени конкретизации смысла тех или иных понятий и отношений между ними, использование KG в учебном процессе позволяет облегчить обучение студентов общим принципам проектирования онтологий и их использованию в практических задачах.

1. Онтология предметной области

Для решения поставленной задачи необходимы онтология грамматики ЕЯ и онтология предметной области анализируемого текста.

В системе KG управления базами знаний имеется ***три*** базовых физических конструкции для хранения информации [Кучуганов, 2002]: дерево – концепт – экземпляр (рис. 1) и ***четыре*** семантические категории для наполнения концептов и экземпляров: предмет; процесс; свойство; отношение (рис. 2).

Дерево задает иерархию концептов. Концепт базы знаний (понятие) определяет подмножество экземпляров, у которых значения параметров удовлетворяют данному понятию, т.е. экземпляры являются листьями деревьев (данными).

База знаний представляет собой семейство деревьев:

<Дерево концептов>::=<*Номер уровня*>, <*Концепт*>, <*Список подконцептов*>,

где подконцепт концепта уровня u – это концепт уровня $u-1$.

Рис. 1. Конструктивные элементы базы знаний

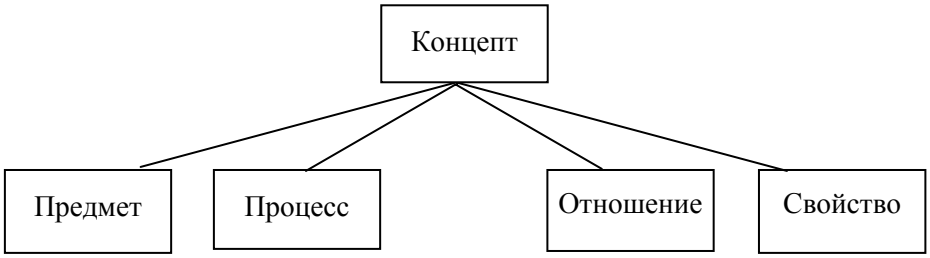

Рис. 2. Виды концептов

Свойства. Обычно в системах для разработки онтологий для каждого понятия задают слоты – свойства, которыми оно обладает. В системе KG свойства описываются в ***разделе знаний о свойствах*** материалов, предметов, процессов, который представляет собой классификационное дерево свойств (атрибутов), сгруппированных по назначению, физическим, физиологическим, психофизическим, социологическим и прочим критериям. Это позволяет унифицировать понятия свойств, использовать их многократно и, самое важное, сопоставлять объекты из разных классификационных деревьев и предметных областей.

Концепт-свойство имеет вид:

<Свойство>*::= <Имя>,[<Комментарий>], <Тип значения>, [<Метод>]*,

где: *Имя* – имя свойства;

Тип значения – непрерывный (real), например, длина, площадь, вес, скорость и т.д.; дискретный (integer), например, количество; денежный; качественный (перечислимый), например, *"малый", "средний", "большой"* и т.п.; текстовый (string); ссылочный ;

Метод – способ вычисления свойства:

* формула для непосредственного вычисления в зависимости от известных атрибутов этого же экземпляра предмета;

* ссылка на сложный метод в разделе процессов.

Предметами здесь считаются материальные и виртуальные объекты: стол, автомобиль, робот, человек (когда он является предметом исследования), документ и т.п. Сюда же входят все виды отчетов от простых до самых сложных. Предметы имеют состав (детали) и схему соединения, т.е. описываются геометрическими моделями: кинематическая схема, чертеж, карта, изображение, 3D геометрическая модель, экранная форма:

<Предмет> *::= <Имя>, <Комментарий>, <Список атрибутов>, [<Список компонентов>, [<Список отношений между компонентами>]], [<Геометрическая модель>].*

*<**Атрибут**> ::= [<Имя>, <Экземпляр-свойство>, <Диапазон изменения значений>, <Ожидаемое значение>], [<Единица измерения>], [<Экземпляр-метод>].*

При конструировании концепта-предмета список его атрибутов набирается из ранее созданных концептов-свойств.

Список компонентов – это дерево входимости деталей в изделие. Во время конструирования концепта-предмета задается типовой состав, которому будут удовлетворять все его экземпляры (хотя там тоже допускаются исключения).

Для определения состава продукта достаточно перечислить все входящие в него подсборки (концепты или экземпляры) уровня *и - 1*, где *и* – текущий уровень сборки. Соответственно, в продуктах уровня *и – 1* перечисляются подсборки или детали уровня *и – 2.* Тогда полное дерево состава сгенерируется автоматически.

Список отношений между компонентами определяет относительное положение, типы соединений, степени свободы и т.п.

Для удобства конструирования концептов в СУБЗ KG выделены два механизма:

- "предок–потомок" – для экономии при описании свойств;
- "сборка–деталь" – для удобства описания составных (многокомпонентных) объектов.

Механизм наследования свойств применяется *к концептам* как средство экономии хранения описаний тех свойств, для которых тип, диапазон, метод вычисления не меняются на младших уровнях иерархии в дереве классов (концептов).

Механизм заимствования компонентов состава применяется к конкретным *образцам* продукта для заимствования тех или иных деталей или подсборок из ранее сконструированных изделий. Они находятся в других деревьях классов и в новой сборке для них пишутся *только отличающиеся параметры*.

Обычно в инженерии знаний этот механизм называют множественным наследованием, применяя к нему тот же инструмент родо-видового наследования, хотя очевидно, что смысл конструирования (агрегирования) совершенно другой – это, как бы, искусственная операция, в отличие от первой – естественной.

Другим принципиальным отличием заимствования компонентов является наличие инструмента параметрического конструирования новых деталей из старых, когда в допустимых пределах меняются те или иные технические характеристики (например, при комплектации персонального компьютера по индивидуальным потребностям клиента) или геометрические (например, размеры), но топология и структура остаются прежними.

Процесс отображает в базе знаний деятельность коллектива исполнителей в некотором отрезке времени и пространства. В общем случае, процессы имеют состав и алгоритм, т.е. их описание содержит вычислительные модели, которые подразделяются на три вида:

- поиск фактов и вычисление свойств;
- анализ ситуаций и исчисление (выявление) отношений;
- действия и работы.

Концепт процесса в общем случае имеет следующий вид:

<Процесс> *::= <Имя>, <Комментарий>, <Список атрибутов>, [<Вычислительная модель>], <Список входных концептов>, <Список выходных концептов>, [<Состав подпроцессов>, <Список отношений подпроцессов>].*

Именно деятельностный процесс порождает множество связей между объектами, поэтому характерной особенностью списка атрибутов процесса является то, что он, помимо общих характеристик содержит атрибуты, описывающие ***роли участников действия***:

- агент (исполнитель);
- бенефициант – заказчик, в чьих интересах выполняется действие, работа;
- реципиент – приемник действия (например, *"Вася дает яблоко Кате"*, реципиент – *Катя*);
- предмет воздействия: исходный/результирующий (в приведенном примере – *яблоко*);
- сцена действия;
- инструмент;
- коагент (соисполнитель) и т.д.

Вычислительные модели задаются с помощью встроенных функций и процедур, внешних программ, а также *схем*, в тех случаях, когда процессы структурированы. При комплексировании процессов, также как и предметов, применяется механизм заимствования и параметризации компонентов состава из разных классов.

Вычислительная модель описывает конкретный алгоритм выполнения некоторой работы. Так, для поиска фактов используются запросы с параметрами.

Процессы анализа ситуаций и выявления отношений устанавливают факты взаимоотношений между двумя или несколькими объектами. В частности, процессы логического вывода, основанные на правилах *ЕСЛИ – ТО*, в качестве предусловий могут содержать достаточно сложные встроенные процедуры и функции анализа фактов.

Действия и работы представляются планами, схемами, графиками.

Списки имен входных и выходных структур данных ссылаются на соответствующие

концепт-предметы и их экземпляры, в том числе это могут быть ***сцены действия*** или их описания, например, план производственного участка, цеха, схема микрорайона, план квартиры и т.п.

Схема процесса задает порядок выполнения подпроцессов. Глубина вложенности или степень детализации не лимитируются по аналогии со структурным подходом к программированию.

Отношения устанавливают факты наличия разнообразных взаимосвязей между объектами. Отношения – это наиболее динамичный раздел базы знаний в том смысле, что они постоянно меняются в ходе осуществления какой-либо деятельности. Поскольку в любой системе поддержки принятия решений факты наличия тех или иных отношений между объектами необходимы, как правило, лишь на этапах анализа и планирования, поэтому при описании предметной области и ее бизнес-процессов имеет смысл определить необходимые виды отношений и способы их исчисления (выявления), а не хранить постоянно сами эти факты. Тогда методы исчисления отношений можно включать в состав процессов (предикатов) анализа ситуаций и принятия решений.

В системе KG предопределены ***четыре*** семантически не пересекающихся категории отношений:

- ***сравнение/сопоставление*** – это сравнение значений свойств (предметов, процессов, отношений) и сопоставление графов (предметов/процессов);
- ***вхождение*** (в множество, класс, экземпляр);
- ***деятельностные*** (ролевые, причинно-следственные, вычислительные) отношения, извлекаемые из экземпляров процессов/задач на основе их концептов;
- ***коммуникации*** (толерантности) – личное отношение к другому субъекту или, в общем случае, к объекту, входящему в другую категорию (предмет, процесс, отношение).

Рис. 3. Семантические категории СУБЗ KG

<Концепт-отношение> *::= <Имя>, [<Комментарий>], <Список атрибутов>], <Имя с уточнителями объекта 1>, < Имя с уточнителями объекта 2>, <Тип значения>, <Ожидаемое значение>, [<Единица измерения>], <Экземпляр-метод>.*

Таким образом, описанная модель знаний позволяет анализировать достаточно широкий спектр отношений между объектами предметной области и бизнес-процессов и, тем самым, уменьшить влияние субъективизма при разработке онтологий. На рис. 3 показаны не пересекающиеся семантические категории концептов СУБЗ KG.

2. Модели элементов текста в среде базы знаний о предметной области

Для решения задач семантического анализа текста в среде базы знаний о предметной области необходимы **[Кучуганов, 2005]**:

1. Предметные онтологии по тематике текстов.
2. Алфавитный словарь основ слов.
3. Справочник "Части речи – Части слова – Схемы сборки".
4. Дерево классов "Правила согласования морфем".
5. Дерево свойств элементов текста (морфологические признаки, синтаксические, семантические).
6. Дерево классов "Правила синтаксического анализа".
7. Дерево правил построения модели сюжета.

Дерево знаний о предметной области содержит концепты (понятия) предметов, процессов, свойств, отношений, а также экземпляры (примеры) этих объектов. Именем каждого концепта является слово или выражение на естественном языке. Перевод слова на другой естественный язык осуществляется с помощью ссылки на соответствующий словарь.

Схема сборки слов с точки зрения лингвистов задает (определяет) базовое правило сборки некоторого класса слов, в то время как множество правил согласования определяется исключениями из этого правила. Типовая схема сборки выглядит так:

[<приставка>[<приставка>[<приставка>]]]<основа>[<суффикс>
[<суффикс>]][<окончание>]

Из этой схемы те или иные детали отбрасываются с помощью правил. На рис. 4 показан пример морфологической схемы слова.

В процессе синтаксического анализа текста на естественном языке осуществляется разбивка сложных предложений на простые, определение свойств и отношений объектов сюжета, а также ролей участников процессов. При этом результаты морфологического анализа слов уточняются с целью устранения противоречий.

Словоформы в сочетании с синтаксическими правилами позволяют определить свойства описываемых в тексте предметов и процессов, их состав, отношения между объектами, порядок сборки сцен и процессов, динамику развития сюжета, поскольку схема сборки слова, правила согласования частей слова и слов в предложении дают смысловые характеристики, необходимые для этого.

uneatable		=	un		+	eat		+	able	
Part of Speech	adjective		Morpheme	prefix		Morpheme	root		Morpheme	suffix
Comp Degree	positive		Part of Speech	N/V		Part of Speech	verb		Part of Speech	verb
			Value	opposite		Value	{EAT}		Value	able to be done

Рис. 4. Пример морфологической схемы слова

Сложноподчиненные предложения обычно возникают, когда автор хочет раскрыть (описать) какие-либо характеристики участников или обстоятельств действия, которые в ближай-

шем тексте не становятся главными, или чтобы различить два похожих объекта (например, "человек, который шел позади, сказал…").

Рассмотрим пример синтаксического анализа текста [Уайт Э.Б. Стюарт Литл]:

(1) The Littles liked to play ping-pong, but the little balls always rolled under chairs, sofas, and radiators. (2) The players had to stop playing and begin to look for the balls. (3) The Littles had a mouse, which was named Stuart. (4) Stuart found them under chairs and hot radiators and pushed them with all his might. (5) The Littles had a piano in their dinning-room, and Mrs. Little liked to play it in the evening. (6) It was a good piano, but one of the keys stuck sometimes, and did not work properly. (7) Mrs. Little said: "It's all because of the bad weather." (8) Mrs. Little's son George always got very angry when he played the piano and the key stuck.

На рис. 5 показана синтаксическая модель последнего предложения.

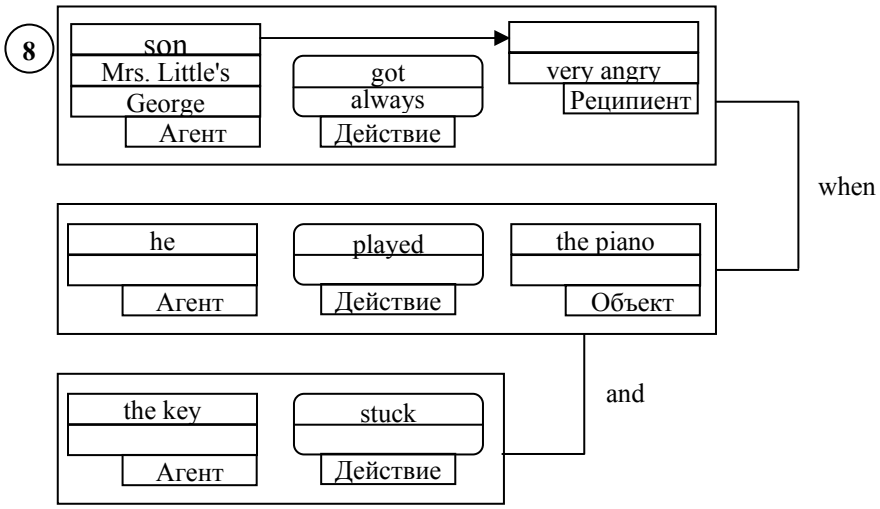

Рис. 5. Синтаксическая модель сложного предложения

В результате синтаксического анализа строится множество моделей простых предложений, объекты которых (предметы и процессы) связаны между собой типовыми схемами процессов базы знаний о предметной области. По мере прочтения текста информация об объектах конкретизируется и дополняется новыми характеристиками, обстоятельствами, отношениями.

Правила синтаксического анализа и правила пополнения модели сюжета имеют продукционный вид "посылка ⇒ следствие". В таком виде они хорошо ложатся в раздел "Методы" предметной области "Грамматика", где множество экземпляров правил сгруппировано по решаемым задачам.

Задача семантического анализа разбивается на 5 этапов:

1. Первичный анализ.

Первичный семантический анализ совмещается с этапом синтаксического анализа, чтобы повысить наглядность визуальных синтаксических моделей. Он осуществляется на основе правил грамматики языка и исключений их них. Например:

∗ *liked to play*: глагол *liked* – это атрибут семьи Литтлов, показывающий их *отношение* к настольному теннису, но этим же словом указывается время существования отношения, поэтому *'liked'* трудно оторвать от основного глагола;

∗ *there is a book*: глагол *to be* выражает факт существования предмета и время существования, но не является действием;

∗ *he had a book*: глагол *to have* здесь показывает факт существования, время и *отношение* собственности;

∗ *to stop, to begin, to end, to quicken, speed up*: выражают действия над действиями или их временные фазы.

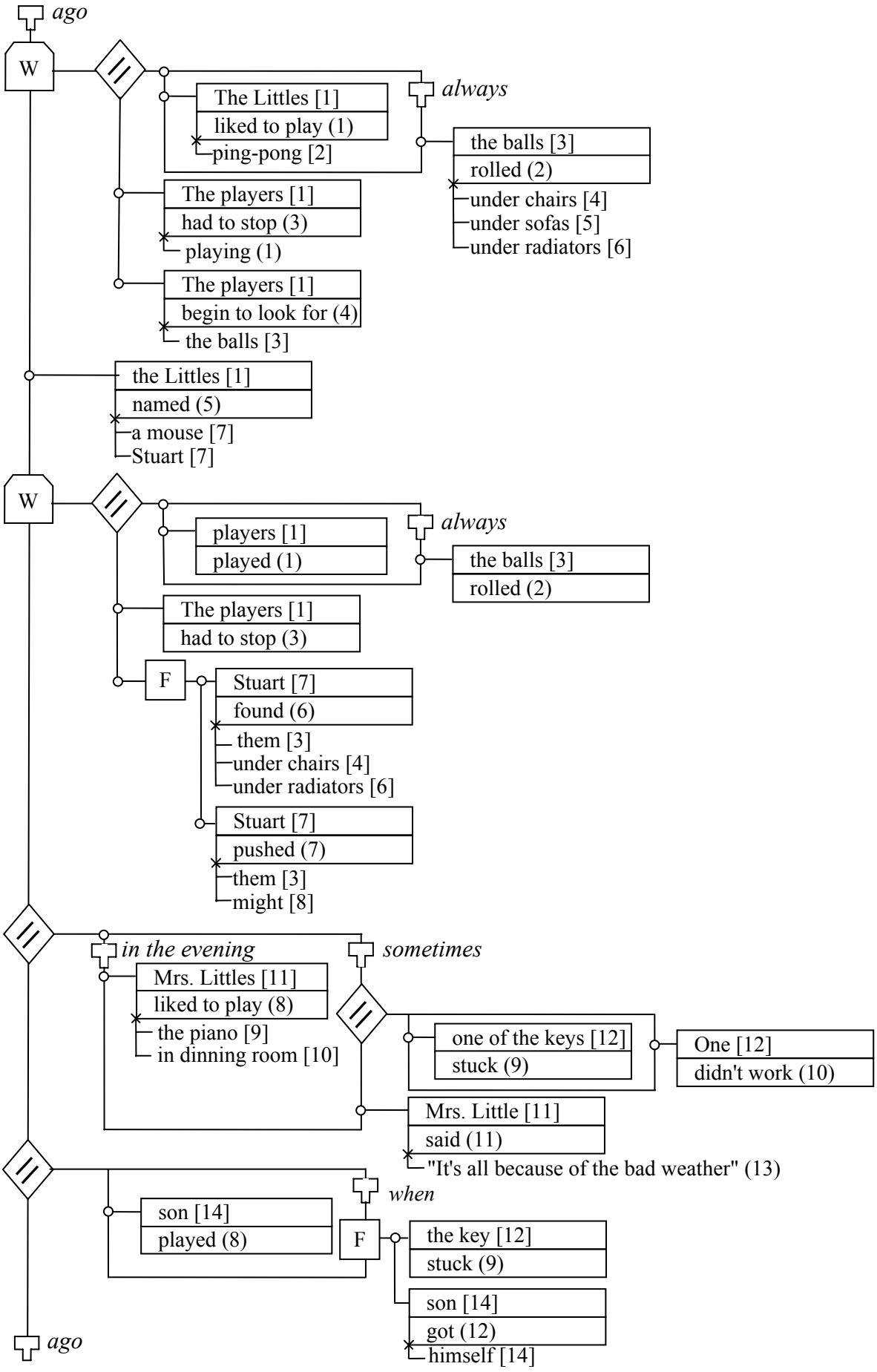

Рис. 6. Пример схемы процессов

На этом же этапе выделяются устойчивые выражения, гиперболы и т.п. и заменяются моделями из словаря.

2. Минимизация количества объектов, т.е. распознавание различных обозначений одного и того же предмета.

3. Распознавание процессов по обозначению, участникам, времени.

4. Синтез схемы процессов текста (рис. 6). Здесь: *F* – последовательность подпроцессов (*Follow*); *R* – циклически с постусловием выполняемые подпроцессы (*Repeat*); *W* – циклически с предусловием выполняемые подпроцессы (*While*); *C* – выбор одного из нескольких подпроцессов (*Case*); // – параллельно выполняемые действия (*Parallel*); ⍗ – отметка времени; ○ – простой (неделимый) процесс (*Simple*); * – перед и после имени процесса показывают, что здесь могут быть выведены имена входных/выходных структур данных (участников).

Отметки времени повышают наглядность схемы, что способствует более качественному принятию решений пользователем, особенно, когда они ставятся в терминах лингвистических переменных.

5. Установление связей с базой знаний о предметной области текста. Множество классов "Части речи" связано ссылками на соответствующие деревья концептов предметов, процессов, свойств, отношений в разделе знаний о предметной области текста. Имена концептов могут быть основами слов или производными конструкциями от них. Следовательно, если имеется словарь, построенный из имен концептов, экземпляров, а также значений перечисляемого типа, то слова анализируемого текста легко связываются с элементами предметной области. Тогда для игры в "вопросы и ответы" достаточно подключить библиотеку функций выявления отношений (см. п.1).

Последовательность моделей простых предложений текста и результирующая визуальная модель сюжета позволяют реализовать обратную связь "воздействие на модель – реакция в тексте", благодаря чему можно в интерактивном режиме отлаживать процессы анализа текстов и доказательства объективности (однозначности) истолкования текстов на естественных языках.

Заключение

Для изображения морфологических, синтаксических, семантических схем мы старались максимально использовать общепринятые символы схем программ и схем бизнес-процессов. На наш взгляд, предложенный подход будет весьма полезен при разработке автоматизированных систем для изучения иностранных языков.

Дальнейшее повышение возможностей визуального моделирования текстов связано с автоматическим синтезом движений по высокоуровневой модели описания требуемых действий (сценарию анимации) для создания трехмерной компьютерной анимации. На рисунке 7 показан пример синтеза сложного движения по сценарию, выполненный в системе MotSyn [Кучуганов и др., 2003]. Система MotSyn – это экспериментальная версия, предназначенная для обучения студентов технологиям трехмерной компьютерной анимации персонажа, обеспечивающая:

- создание и редактирование движений по технологии «ключевых кадров» (key framing);
- создание новых движений путем модификации ранее созданных движений на основе пространственно-временных ограничений и требований, заданных с помощью логических правил;
- создание нового движения путем комбинирования двух движений (последовательного или одновременного выполнения);
- синтез сложного движения персонажа по сценарию – высокоуровневому описанию требуемых действий.

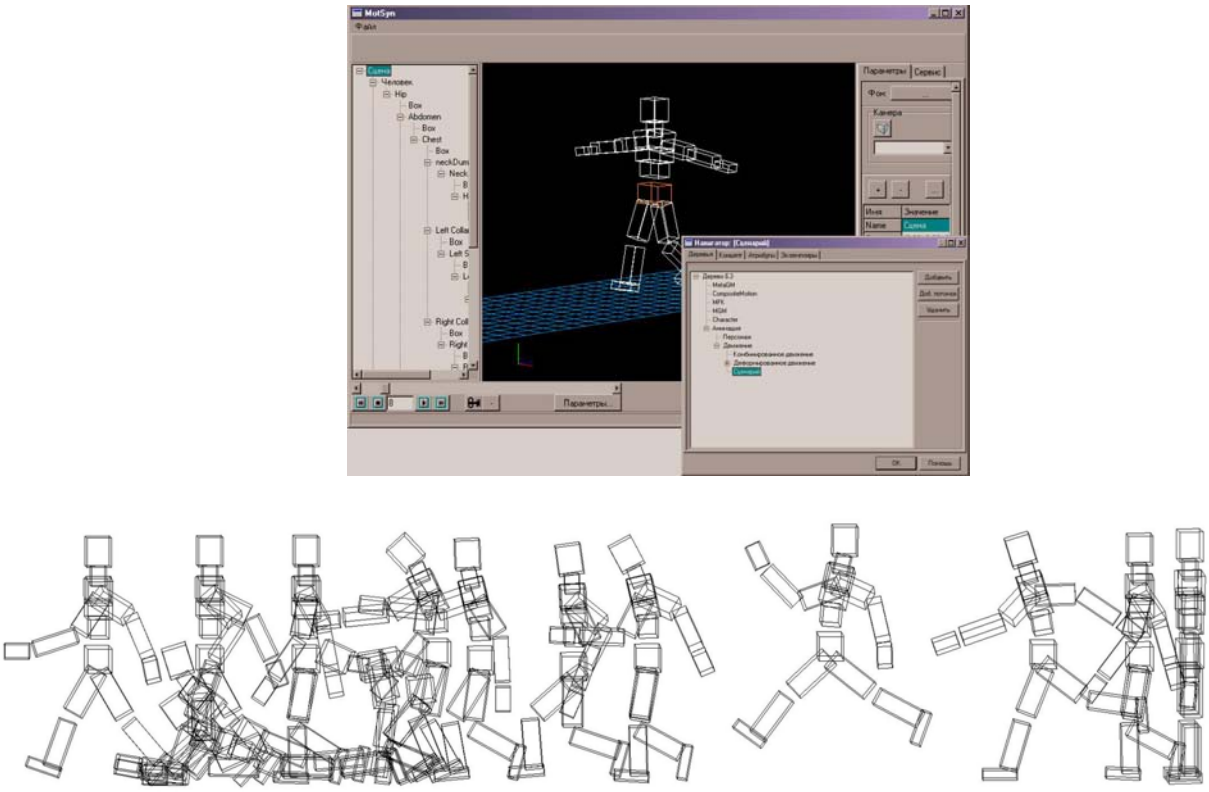

Рис. 7. Пример синтеза сложного движения по сценарию

Система MotSyn 1.0 интегрируется в состав системы KG, что позволяет использовать базу знаний для представления персонажей и их движений, а также для логического вывода, применяемого при синтезе сложного движения.

Литература

[Гаврилова, 2000] Гаврилова Т.А. и др. Базы знаний интеллектуальных систем // Учебник для вузов. – СПб. Изд-во "Питер", 2000.

[Кучуганов, 2002] Кучуганов В.Н. Семантика графической информации. Известия ТРТУ. Тематич. вып. "Интеллектуальные САПР". Материалы междунар. научн.-техн. конф. "Интеллектуальные САПР". Таганрог: Изд-во ТРТУ, 2002, №3(26). С. 157-166.

[Кучуганов и др., 2001] Кучуганов В.Н., Габдрахманов И.Н. Система визуального проектирования баз знаний. – Информ. технологии в инновационных проектах: Труды III междунар. науч.-техн. конф. – Ижевск, 2001. С. 140-143.

[Попов, 1987] Попов Э.В. Экспертные системы. Решение неформализованных задач в диалоге с ЭВМ. М.: Наука, 1987. – 288 с.

[Кучуганов, 2005] Кучуганов В.Н. Визуальное моделирование текстов // Труды Междунар. научно-технич. конференций "Интеллектуальные системы" (AIS'05)" и "Интеллектуальные САПР" (CAD-2005). – М.: ФИЗМАТЛИТ, 2005. – Т. 4. С. 104-114.

[Кучуганов и др., 2003] Кучуганов В.Н., Семакин М.М.. Синтез движений в задачах трехмерной компьютерной анимации персонажа / Материалы Международной научно-технической конференции IEEE AIS’03, CAD-2003 (3-10 сентября 2003г.).- Том II. Информационные технологии. − Москва: Изд-во Физматлит, 2003. С. 229–235.