ПРЕДСТАВЛЕНИЕ НЕКОТОРЫХ АСПЕКТОВ РУССКОГО СЛОВООБРАЗОВАНИЯ В ОБЪЕКТНОЙ МОДЕЛИ МНОГОФУНКЦИОНАЛЬНЫХ ЭЛЕКТРОННЫХ СЛОВАРЕЙ

REPRESENTATION OF SOME ASPECTS OF RUSSIAN WORD FORMATION IN MULTIFUNCTIONAL COMPUTER DICTIONARY OBJECT MODEL

О.О. Ивличева

РГГУ, Москва

ivlitcheva@rambler.ru

В статье рассматривается дальнейшее применение предложенной ранее (Диалог'2003) объектной модели для разработки и ведения электронных словарей. На нескольких примерах демонстрируются возможности модели для представления в словаре данных, связанных со словообразованием в русском языке.

В работах [1, 2] был предложен новый подход к программным реализациям многофункциональных электронных словарей (под электронным словарем мы понимаем программу, способную хранить и обрабатывать словарные данные). Эта работа велась в рамках общего проекта по объектному моделированию в лингвистике и представлению текстов на естественном языке, разрабатываемого в Отделении интеллектуальных систем (в гуманитарной сфере) Института лингвистики РГГУ.

Предлагаемая объектная модель для разработки и ведения электронных словарей обладает более широкой, чем это обычно принято в компьютерной лексикографии, функциональностью. При её разработке мы ориентировались на следующие задачи, выходящие за рамки простых запросов к словарю: разработка новых словарей в ситуации, когда состав и структура словарных статей изначально не до конца прояснены и могут изменяться в процессе работы; неоднократное пополнение словаря из различных источников; построение по универсальному словарю новых, специфицированных пользователем словарей.

Отличительными свойствами предлагаемой объектной модели являются: локальность представления данных; отражение свойств как отдельных объектов, так и их частей; ориентированность на синтез; возможность оперативного анализа структуры и состава словаря.

Входами словаря на уровне представления информации являются объекты (в смысле объектно-ориентированного программирования), обладающие свойствами и поведением. В виде объектов, объединенных в многоссылочную структуру, представляются языковые единицы, информация о них и отношения между ними. За счет такой архитектуры достигается многофункциональность, которая позволяет на основании одного и того же формата представления словарной информации решать различные задачи, как лексикографические, так и связанные с классическими проблемами обработки текстов на естественном языке.

Словари, разработанные в такой объектной среде, являются фактически результатом обработки объектной модели. Объектная модель состоит из так называемых объектов представления (R-объектов – от *representation*) с соответствующим наполнением и связей между ними, соединяющих объекты в сеть. Хотя общая структура R-объектов схожа (в смысле набора атрибутов, которыми они обладают), разные R-объекты представляют разные наборы свойств, относящихся к предметной области. Это достигается за счет того, что хранение "предметных" свойств сущности, которую представляет объект, организовано в форме списка произвольной длины. С точки зрения конечного пользователя, объекты становятся гибко настраиваемыми – без жесткой, однозначно заданной с самого начала, структуры данных.

Все свойства, представляемые R-объектами, можно разделить на две непересекающиеся группы: аддитивные и внутренние. Внутренние свойства характеризуют только сам объект, а аддитивные наследуются старшим объектом – контейнером, включающим в себя объект-хозяин свойства (подробнее о структуре и базовом поведении объектов см. [1]).

Каждый объект соответствует осмысленной лингвистической сущности. В целом R-объекты организованы в сеть, иерархия которой соответствует естественной иерархии самих лингвистических единиц. Старшинство объектов в рамках этой сети отражает вхождение более простых единиц в более сложные. Входом словаря может быть любая значимая лингвистическая единица, при этом более сложные получаются в процессе синтеза из более элементарных. Таким образом, составные единицы хранятся в словаре не самостоятельно, а в виде своеобразных «формул», по которым их можно получить.

Каждому входу словаря сопоставлен его экстенсионал, или объем, т.е. множество текстов, которое он представляет. Вычисление экстенсионала осуществляется по запросу,

обращенному к объекту. Запрос, на который можно накладывать различные условия, является частью поведения объектов и аналогичен операции селекции в реляционных базах данных. Вычисление запроса происходит в результате вывода, то есть рекурсивного вычисления экстенсионалов объектов, подчиненных объекту, к которому запрос применен, и применения к ним операции соединения текстов. Эта операция комбинирует наборы свойств и тексты, хранящиеся в объектах – её аргументах.

Свойства составляющих словарь R-объектов также реализованы в виде объектов, обладающих собственным поведением. Это поведение, или обработчик свойств хранящихся в словаре лингвистических единиц, может быть задано пользователем с помощью алгоритма [4, 2] – совокупности правил, организованных в сеть, в которой они являются вершинами. Связи между узлами сети соответствуют переходам от правила к правилу в процессе вычисления алгоритма. В этом смысле правило может иметь не более двух дочерних узлов, причем при выполнении условия правила осуществляется переход к одному из них, а при невыполнении – к другому. Использование алгоритмов стало удобным способом упростить использовавшиеся сначала для моделирования различных сложных случаев конструкции из объектов с более простым поведением, эквивалентные по выразительности, но достаточно громоздкие. При этом добавление алгоритмов не повлияло на представление более простых по своему поведению лингвистических единиц. Описание поведения самих единиц становится при этом более компактным и технологичным.Покажем, как поведение свойств, описываемое алгоритмом, используется для моделирования различных лингвистических единиц. Поскольку минимальными единицами с ненулевым означающим, представленными в словаре, являются морфемы, было бы желательно сделать словарь по возможности естественным отражением морфемики русского языка, учитывающим знания о типах, структуре и поведении морфем. Представление в объектной модели морфемики как части грамматики, охватывающей те аспекты словообразования и морфологии, которые связаны с аффиксами и выражаемыми с их помощью словообразовательными и грамматическими значениями, является необходимым этапом в построении адекватной модели лексики языка.

Покажем один из способов отображения свойств и поведения морфем, состоящих из нескольких алломорфов, на примере суффикса "-ин-" («господин», «горожанин»), который «пропадает» в словоформах множественного числа («господа», «горожане»).

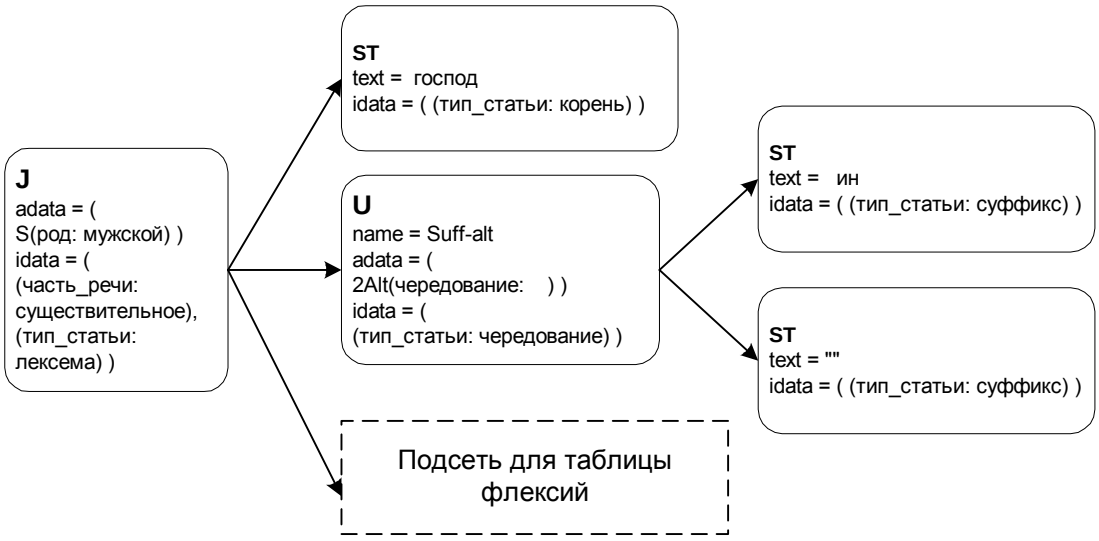

Рис. 1 Представление лексемы «господин»

На рис.1 показано представление лексемы «господин» в нотации, предложенной в работе [1] . В целях сокращения и наглядности объекты здесь изображены упрощенно (присутствуют только существенные для данного примера поля данных). Объемом корневого объекта в структуре, обозначенной как «Подсеть для таблицы флексий» (т.е. множеством лингвистических единиц, которые объект перечисляет при запросе к нему), будет множество из 12 окончаний с признаками числа и падежа (подробнее о представлении таблиц флексий см. [1]). Объем для объекта Union с именем `Suff-alt` – это множество

{ST[text=«ин», adata=(2Alt(чередование:)),idata=…],
ST[text=«», adata=(2Alt (чередование:)), idata=…]},

т.е. при запросе к нему будут порождены копии его «детей», наделенные свойством чередоваться. В соединении на уровне лексемы будут рассматриваться тройки из объекта-корня в качестве первой компоненты, этих двух объектов в качестве второй компоненты и каждой из флексий в качестве третьей.

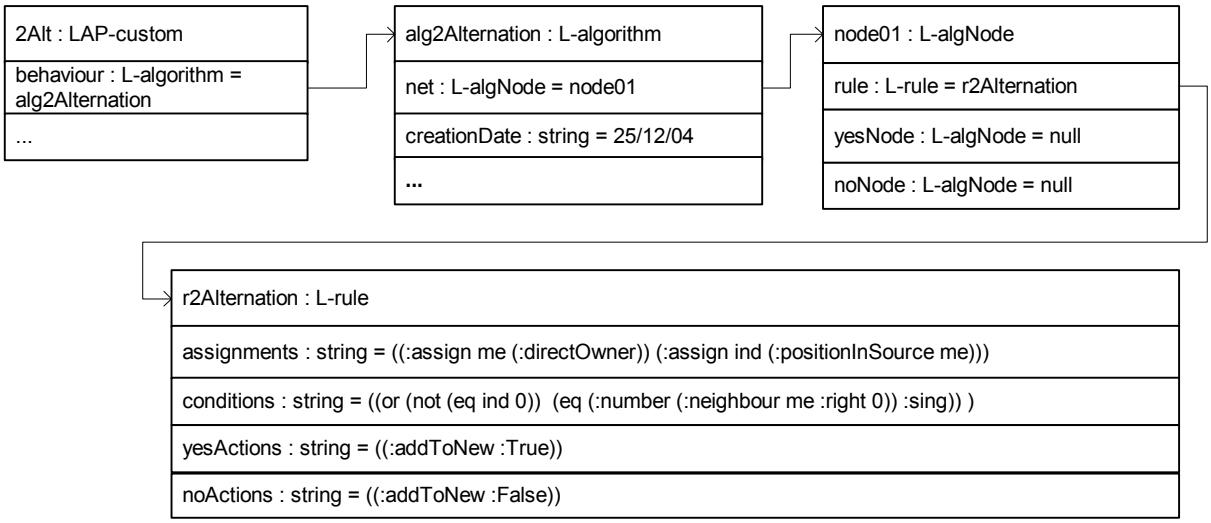

Рис. 2 Объект-обработчик свойства, задающего чередование, подобное случаю ин/Ø

Поведение обработчика свойства «чередоваться подобным случаю *ин/Ø* образом» задано алгоритмом на рис. 2. Здесь приведена подробная диаграмма структуры этого объекта так, как она строится в графическом конструкторе структуры объектов, описанном в работе [3] (мы лишь не показываем «служебные» слоты у некоторых объектов, несущественные в рассматриваемом примере).

Сеть этого алгоритма содержит всего один узел, так как поведение описывается единственным правилом (подробнее об этих объектах см. в [4]). При его применении вначале выполняются присваивания его локальным переменным `me` и `ind`. Форма `(:directOwner)` определяет объект – непосредственный владелец выполняемого свойства, т.е. значением переменной `me` будет присутствующий в рассматриваемой паре на соединение один из двух объектов, порожденных объектом Union с именем `Suff-alt`. Переменная `ind` получит в качестве значения номер позиции, на которой такой объект находится в объеме исходного для него объекта `Suff-alt`, этот индекс определяется формой `(:positionInSource me)`. В условии правила проверяется, что если текущим является первый элемент чередования (в данном примере – суффикс *-ин-*), то ему должно соответствовать единственное число. В противном случае выполняется форма `(:addToNew :False)`, в которой ключевое слово `:False` запрещает порождение нового объекта – соединения анализируемой тройки из корня, суффикса и флексии. Аргумент `:True` такой формы наоборот, разрешает порождение, но он, по существу, является dummy-значением («заглушкой»). Он заменяет форму, добавляющую новые аддитивные свойства в порожденный объект в качестве результата обработки чередования. Это, в частности, означает, что само свойство чередования далее «всплывать» (т.е. передаваться в качестве аддитивного свойства объектам, построенным с использованием данного порожденного) уже не будет. Мы могли бы оставить его в порожденных объектах словоизменения лексемы, например, добавив ему в качестве значения порядковый номер суффикса при чередовании (`ind`), но тогда мы уже не обошлись бы единственным правилом в алгоритме – нужно было бы следить за тем, когда свойство должно быть обработано, а когда уже нет.

Чередование суффиксов, рассмотренное выше, – не единственный пример чередования, где признаком, определяющим выбор альтернанта, является число словоформы. В качестве примера можно привести почти все существительные с суффиксами *-онок-/-ат-*, *-ёнок-/-ят-* («ежонок», «слонёнок» и т.д.). Похожим поведением обладают языковые единицы ("формативы") типа *Ø/ес* в словах "небо"/"небеса", "чудо"/"чудеса" и т.д. (такие единицы, согласно некоторым исследователям, не могут быть отнесены к морфемам в силу своей "асемантичности"). При предлагаемом представлении такое чередование будет задаваться тем же самым свойством, которое использовалось выше в описании суффикса «-ин-». Причем согласно принципу локальности определяющий поведение этого свойства алгоритм будет существовать в модели в единственном экземпляре. Его локальное изменение приведет к изменению обработки подобного чередования во всей модели.

Как видно из примера, в предлагаемой объектной модели может содержаться отражение свойств и поведения отдельной произвольной морфемы. С другой стороны, построенная таким образом объектная структура данных моделирует словообразование в его статическом аспекте (по Л.В. Щербе), т.е. отражает, "как сделаны слова". Если парадигма слова получается в результате вывода на объектной модели, то его морфемная структура фиксируется непосредственно самой моделью.

Таким образом, словарь с предлагаемой структурой данных можно считать первым шагом в разработке модели, отражающей словообразование в русском языке. Разработанные механизмы могут быть использованы для моделирования не только морфем, но и других теоретических понятий словообразования. Если же потребуются новые структуры данных, что вероятно ввиду большого количества и разнообразия единиц системы словообразования (от аффиксов до словообразовательных категорий и других объединений производных), их добавление не повлияет на работоспособность уже описанного ядра модели.

Выразительная способность разработанных объектов в силу их гибкости и отсутствия у них фиксированного набора свойств достаточно высока. Например, объектная модель в существующем сейчас виде отражает для корневых морфем состав образованного ими словообразовательного гнезда, но не систему отношений слов, входящих в него. Однако эта система в большинстве случаев может быть представлена в виде надстроенной дополнительной модели из уже существующих объектов. К примеру, любая древовидная структура может быть промоделирована с помощью объектов класса Union.

Рассматривая свойства предлагаемой объектной модели, можно отметить, что более естественно в ней отражаются процессы деривации, связанные с "наращением" структуры слова – аффиксальное словопроизводство и словосложение. С другой стороны, конверсия в основном может быть промоделирована с помощью тех же механизмов, что и омонимия. Введение дополнительных структур может потребоваться только для представления аббревиации, если мы захотим отразить её в виде процесса получения новых слов, а не только в виде фиксации родственных отношений между исходными и производными словами.

Представляется также, что при соответствующем наполнении словарь, основанный на такой объектной модели, мог бы стать инструментом изучения дистрибуции морфем. Запросы к модели и механизм коллекций (подробнее о коллекциях см. [1]) позволяют выделять и сохранять для каждой морфемы множество её возможных окружений.

В заключение можно сказать, что целью данной работы было получение удобной и открытой базовой структуры данных с возможностью легкого встраивания в нее интеллектуальных компонент. Эта структура имеет определенный, специфицированный интерфейс и может использоваться в дальнейших работах по различным словарям, по строению словарей, а также по моделированию лексики. По существу вышеописанные примеры моделирования в предлагаемой объектной модели являются имитационным моделированием различных понятий теории словообразования.

Список литературы:

1) Ивличева О.О., Епифанов М.Е., Лахути Д.Г. Объектная модель многофункциональных словарей, основанная на синтезе лингвистических единиц // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003. – М.: Наука 2003, с. 223-231.
2) Ивличева О.О. Эксперименты с представлением некоторых сложно устроенных составных лингвистических единиц в объектной модели многофункциональных электронных словарей // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. Т.2. – М.: Физматлит, 2004, с. 525-534.
3) Ершова Е.С., Епифанов М.Е. Графический конструктор структуры объектов как интерфейс инструментальной объектной среды // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. Т.2. – М.: Физматлит, 2004, с. 525-534.
4) Баталина А.М., Епифанов М.Е., Ивличева О.О., Кобзарева Т.Ю., Лахути Д.Г. Инструментальная среда для экспериментов с алгоритмами поверхностно-синтаксического анализа // Труды международной конференции Диалог'2004.
5) Кузнецова А.И., Ефремова Т.Ф. Словарь морфем русского языка. - М.: Русский язык, 1986.
6) Тихонов А.Н. Словообразовательный словарь русского языка. - М., 1985.
7) Уорт Д.С Русский словообразовательный словарь. Введение // Новое в зарубежной лингвистике. Вып.14 : Проблемы и методы лексикографии. - М.: Прогресс, 1983.
8) Янко-Триницкая Н.А. Словообразование в современном русском языке. – М.: Издательство «Индрик», 2001.