УДК 519.62

Г.Ф. Дюбко, Е.С. Карлин

*Харьковский национальный университет радиоэлектроники, Харьков*

# МОДЕЛИ СЕМАНТИКИ ПРОСТЫХ ПРЕДЛОЖЕНИЙ РУССКОГО ЯЗЫКА И ИХ ПРИЛОЖЕНИЕ В СИНТАКСИЧЕСКОМ АНАЛИЗЕ

*Основной идеей данной модели является проектирование анализатора, который производит синтаксико-семантический анализ простого предложения русского языка, определяя семантические функции предложения и используя Л-словарь. Семантическая функция – это определенный смысл, который ставится каждому словосочетанию (группа связанных между собой по смыслу и грамматически слов) предложения. Словосочетание в свою очередь формируются с помощью поверхностного смысла входящих в него слов, который определяется толковым словарем.*

***Ключевые слова:*** *семантика, синтаксис, морфология, база данных, анализ, падеж, предложение, естественный язык, λ-словарь, семантическая функция, периферия.*

## Введение

Проблема интеллектуального анализа и понимания текстов на естественном языке появилась одновременно с созданием компьютеров. Однако до настоящего времени достичь полного успеха в её решении не удалось. Методы для решения частных задач разработаны в большом количестве, но количество так и не перешло в качество.

Накопление огромного количества информации в текстовом виде делает актуальным создание программ, ориентированных на ее обработку. Диапазон таких программ весьма широк: системы автоматического перевода, естественно-языковые интерфейсы, системы автоматического реферирования. Несмотря на широту диапазона решаемых задач, качество работы всех этих систем прямо зависит от уровня формализации естественного языка, который может предложить современная теория искусственного интеллекта. Большое количество современных работ в области искусственного интеллекта направлены на разработку моделей семантики, которые позволят сделать качественный скачок в семантической интерпретации текстов и улучшить результаты практической работы систем обработки текстовой информации.

Естественный язык (ЕЯ) обладает многими свойствами, необходимыми, для представления знаний. Но для того чтобы использовать его как средство представления знаний в ЭВМ, необходимо решить сложнейшую проблему содержательного (семантического) анализа текстов на ЕЯ. Вопросам формализации отдельных сторон ЕЯ посвящено немало работ как у нас в стране, так и за рубежом. Большинство работ по формализации семантики языка состоит в разработке формальных моделей и языков их представления, которые должны позволять одинаковое формальное каноническое представление для различных естественно-языковых конструкций (ЕЯК), имеющих одинаковый смысл.

## Постановка задачи

Для создания естественно-языковых интерфейсов компьютерных систем необходимо иметь:

- язык представления знаний в системе для описания смыслов слов естественного языка, так называемую семантический язык, систему отношений между понятиями естественного языка в виде, удобном для использования в компьютерном анализе;

- словари естественного языка, в которых бы этот язык был представлен в полном объеме;

- программу (лингвистический процессор), которая позволяла бы выполнять операции выделения смыслов из текстов на естественном языке (анализ) и обратные операции - создание смыслов и представление их на естественном языке (синтез).

В данной работе будут рассматриваться такие решения из задач: создание семантического языка и разработка программы, которая позволяла бы выполнять операции выделения смыслов из текстов на естественном языке. Для их решения необходимо:

- изучить существующие современные принципы формализации естественного языка, методы анализа текста на естественном языке и способы построения лингвистических процессоров; существующие подходы к организации морфологических и семантических словарей, словарей словосочетаний, методику пополнения словарей лексикой естественного языка;

- разработать модель предметной области и внутреннюю структуру словарей, что позволяет эффективно использовать сохраненную в них информацию, способ хранения в словарях морфологической и семантической информации;

- определить требования к программной среде, в которой будут разрабатываться словари и выбрать соответствующее программное среды разработки;

- разработать систему, ориентированную на обработку простых предложений русского языка и предоставления информации о семантической корректности предложения.

При разработке системы учитывается тот факт, что семантическая сеть уже разработана и она не может быть изменена в данной системе.

## Синтаксический анализ и формирование семантических функций

Реализация алгоритма начинается с морфологического анализа каждого из слов предложения и определения для него различных морфологических признаков, таких как часть речи и в зависимости от нее – число, род, падеж, спряжение, склонение и т.д. На следующем этапе с помощью этих признаков происходит синтаксический анализ предложения и формируются простые словосочетания. Для них определяются свои семантические функции. Например, словосочетанию «скорый поезд» будет соответствовать семантическая функция

F_прилагательное (Vi(«поезд»), Vj(«скорый»)),

где Vi(«поезд»), Vj(«скорый») – поверхностные смыслы для слов «поезд» и «скорый» соответственно. Далее происходит повторный анализ предложения и формирование сложных словосочетаний, состоящих из более простых. Например, для словосочетания «большой скорый поезд» при первом проходе формируется семантическая функция, рассмотренная выше.

При втором проходе формируется сложная функция

F_прилагательное (Vk(«большой»),
F_прилагательное (Vi(«поезд»), Vj(«скорый»))),

т.е. анализ осуществляется многопроходным процессом. Если после этого процесса набор семантических функций изменился – он запускается опять. Иначе анализ останавливается и на выходе мы получаем набор конечных словосочетаний и семантических функций для них.

Формирование рассмотренных выше семантических функций осуществляется благодаря за ранее сформированным продукциям и управляющей таблице. Рассмотрим более подробно эти понятия.

Каждое слово имеет свои признаки. Продукции формируют из одних семантических функций новые, более сложные, оперируя этими признаками. Продукция имеет следующий вид:

1) <F_генетив>.a1 -> xy;

2)y.part_of_speech == “существительное”, x.part_of_speech == “существительное”, y.case == “родительный”;

3) <F_генетив>.a1 = F_генетив(V(x), V(y)).

Данная продукция показывает, что если мы имеем два слова и часть речи каждого из них – существительное, а также падеж второго слова родительный, то данное словосочетание является семантической функцией F_генетив.

Для выполнения алгоритма используется стек. Каждое слово либо сформированная семантическая функция переносятся в стек и далее возможна свертка двух верхних элементов стека в семантическую функцию. Управляющая таблица необходима именно для управления процессами переноса в стек и свертки. Как пример, ниже приведен фрагмент данной таблицы:

Таблица 1

Фрагмент управляющей таблицы

| | прилагательное | существительное | ... |
|---|---|---|---|
| прилагательное | перенос | перенос, свертка P1 | ... |

Данная таблица представлена двумерным массивом. В нулевом столбце каждой строки находится часть речи слова либо тип семантической функции элементов, которые могут находиться в стеке. В нулевой строке каждого столбца находится часть речи слова либо тип семантической функции элементов, которые могут идти следующими в предложении. На пересечении строк и столбцов определены действия, которые необходимо произвести. Т.е. например, если в вершине стека находится слово, часть речи которого прилагательное и часть речи следующего за ним слова та же, то необходимо произвести операцию переноса в стек для второго элемента. Если же часть речи второго слова – существительное, то также производится его перенос в стек и далее производится свертка двух верхних элементов стека в семантическую функцию по продукции P1, описывающей формирование F_прилагательное. Стек заполняется, пока не закончится предложение, далее содержимое стека переводится в очередь и снова подвергается анализу с помощью управляющей таблицы.

## Семантические примитивы

Значения каждого слова содержится в семантическом словаре и выражено посредством семантических единиц (элементарных смыслов), с помощью которых представлена и сама семантическая структура ЕЯ-конструкций. Простейшей семантической единицей служит семантический примитив, выражающий значение единицы в реальном мире. Лексическая семантика базируется на множестве таких примитивов - элементарных единиц смысла.

Элементарные смыслы могут образовать иерархическую структуру на верхнем уровне которой находятся примитивы ОБЪЕКТ, ДЕЙСТВИЕ, АБСТРАКЦИЯ, ОТНОШЕНИЕ. Каждому из этих примитивов соответствуют примитивы более низкого уровня.

Например, класс семантических примитивов ОТНОШЕНИЕ может содержать такие семантические единицы, как ПРИЧИНА, СЛЕДСТВИЕ, РОДСТВО, ПРИНАДЛЕЖНОСТЬ, и т.п.

Семантически словарная статья (значение слова) описывается следующим образом. Каждому слову α языка сопоставляется семантическая функция $V_i$ (α), где i – номер словарной статьи слова α в семантическом словаре. Значением функции $V_i$ является имя семантического примитива, говорящее о смысле слова α в употребляемом контексте. Семантика словосочетаний выражается семантической функцией F_имя, состоящей из более простых семантических функций: F_имя($X_1$,...,$X_n$), где F_имя - наименование семантической функции, $X_i(1 \le i \le n)$ - семантическая функция $V_i$ (α).

Используя семантические функции и примитивы, соответствующие аргументам этих семантических функций можно вычислять примитивы, являющиеся семантическим значением семантической функции. Если семантическая функция представляет собой словосочетание, то ее значение соответствует значению главного слова в ней.

Таким образом, каждой семантической функции можно поставить в соответствие определенный смысл, выраженный семантическим примитивом.

## Λ-словарь и периферия

Лямбда-исчисление, предложенное А. Черчем, является теоретической основой описания вычислительных процессов, которая не содержит в явном виде понятие ячеек памяти для хранения значений переменных и последовательности вычислений как процесса изменения состояния памяти. А. Черч построил систему, где используются правила преобразований, с помощью которых можно получать из одних функций другие, эквивалентные им. Λ-исчисление позволяет объединение функционального и предикатного подходов, т.е. в λ-формулы можно встраивать формулы исчисления предикатов.

Объединение λ-формулы с предикатными формулами и семантическими примитивами позволяет создать механизм для вычисления словарной статьи глагола и его периферии (употребление глагола в конкретном предложении) исходя из результатов синтаксико-семантического анализа предложения. Для этого с каждой словарной статьей глагола α связывается соответствующим образом построенная λ-формула.

В общем случае такая λ-формула имеет вид:

$$\lambda x_1.\lambda x_2.\lambda x_n.(P1(X1)\text{^}P2(X2)\text{^}\ldots Pm(Xm)\text{^}Q1(X1,\ldots,Xn) \text{^}Qk(X1,\ldots,Xn))@\omega 1\ldots@\omega n \quad (1)$$

где PI(XI), $(1 \le I \le n)$, Qi(X1,…,Xn) $(1 \le i \le m)$ – предикаты;

имя предиката PI – семантический примитив;

предикаты QI – формализуются как предикаты равенства морфологических признаков соответствующих аргументов каким-то вполне определенным значениям;

ω1,…,ωn – периферия, которая определяется результатом синтаксико-семантического анализа глагола α в конкретном предложении, т.е. ω1,…,ωn – это семантические функции для внутренних словосочетаний предложения.

Предикаты Pi и Qj принимают значения «истина», если семантический примитив xi (аргумент Pi) равен Pi, а для Qj выполняются соответствующие равенства. В целом истинность λ-формулы определяется конъюнкцией ее составляющих (предикатов) после выполнения редукции, т.е. подстановки ω1,…,ωn вместо аргументов в (1).

С каждой словарной статьей глагола α связана одна или более формул типа (1). Выбирается та словарная статья, для которой формула типа (1) становится истинной после редукции λ-формулы.

Отметим, что множество {ω1,ω2,…,ωn} в (1) рассматривается как упорядоченное множество, т.е. предполагается фиксированный порядок слов в анализируемом предложении, чего нет в реальной языковой практике. Поэтому в процесс вычисления соответствующей словарной статьи необходимо рассматривать все перестановки множества {ω1,ω2,…,ωn} (с изменением порядка следования словосочетаний), и выбирается та перестановка, которая дает истинный результат.

С каждой словарной статьей связан вполне определенный вид семантической функции с аргументами, обозначенными переменными из (1), в которую вместо переменных подставляются значения <ω1, …, ωn>. Эта семантическая функция выбирается в качестве значения анализируемого предложения.

Формирование λ-формулы типа (1) можно осуществить путем анализа и формализации словарных статей глагола α в толковом словаре. Рассмотрим на конкретных примерах формирование формул для глаголов, обозначающих действия.

Глагол «работать» в толковом словаре Ожегова С.И. имеет 5 словарных статей.

Для «работать» в значении «находиться в действии» λ-формула будет иметь вид:

λx.(ОБЪЕКТ(x))@ω

Это справедливо для предложений «Машина работает», «Завод работает».

Для «работать» в значении «осуществлять деятельность» λ-формула будет выглядеть следующим образом:

λx1.λx2.(ОБЪЕКТ(x1)&ОБЪЕКТ(F_предлог(V («над»),V(x2)))&x1.одушевленность=«одушевленный») @ω1@ω2

Это справедливо для предложений типа «Работать над рукописями».

Для «работать» в значении «иметь постоянное занятие» λ-формула будет выглядеть следующим образом:

λx1.λx2.(ОБЪЕКТ(x1)&СПЕЦИАЛЬНОСТЬ(x2) ^x1.одушевленность=«одушевленный»&x2.падеж=предложный»)

Это справедливо для предложений типа «Работать слесарем», «Работать дворником».

Для глагола «идти» в значении «двигаться, переступая ногами» λ-формула будет выглядеть следующим образом:

λx1.λx2.(ОБЪЕКТ(x1)&МЕРА(x2)&x1.одушевленность=«одушевленный»&x2.падеж=творительный») @ω1@ω2

Это справедливо для предложений типа «Лошадь идет шагом».

Сформированная вышеописанным способом информация содержится в семантическом словаре, названным λ-словарем. Здесь каждой словарной статье конкретного глагола приписана λ-формула типа (1). Эта формула становится истинной, если периферия глагола совпадает с периферией в λ-формуле. Также проверка происходит на уровне словосочетания и для главного слова каждого словосочетания также сравнивается его периферия с периферией в λ-формуле.

## Выводы

Данный анализатор позволяет проводить анализ простых предложений русского языка и прдставлять их в формализованном виде, используя семантические функции и λ-выражения.

Основное отличие данного анализатора от других заключается в использовании λ-исчислений для описания семантики глаголов, выраженной в формальной семантике. На основе интерпретации толкового словаря принимается решение об осмысленности предложения и формирование выходной семантической функции.

В будущем планируется расширение анализатора и возможность обработки сложных предложений, а также его интегрирование в другие естественные языки.

## Список литературы

*1. Любко Г.Ф. Формальная семантика и анализ естественного языка / Г.Ф. Дюбко, Д.В. Преснякова. – Х.: ХНУРЭ, 2006. – 180 с.*

*2. Любко Г.Ф. Некоторые аспекты формализации толкового словаля с применением Л-исчисления / Г.Ф. Дюбко, М.Ю. Дзёх, Д.В. Преснякова. – Х.: ХНУРЭ, 2008. – 144 с.*

*Поступила в редколлегию 1.04.2011*

**Рецензент:** д-р техн. наук, проф. С.Ю. Шабанов-Кушнаренко, Харьковский национальный университет радиоэлектроники, Харьков.

### МОДЕЛІ СЕМАНТИКИ ПРОСТИХ РЕЧЕНЬ РОСІЙСЬКОЇ МОВИ ТА ЇХ ДОДАТОК В СИНТАКСИЧНОМУ АНАЛІЗІ

Г.Ф. Дюбко, Є.С. Карлін

*Основною ідеєю даної моделі є проєктування анализатору, який здійснює синтаксико-семантичний аналіз простого речення російської мови, визначаючи семантичні функції речення і використовуючи Л-словар. Семантична функція – це деякий сенс, що ставиться кожному словосполученню (група зв'язаних між собою по змісту та граматично слів) речення. Словосполучення в свою чергу формується за допомогою поверхневого змісту вхідних в нього слів, який визначається толковим словником.*

***Ключові слова:*** *семантика, синтаксис, морфологія, база даних, аналіз, відмінок, речення, природна мова, λ-словник, семантична функція, періферія.*

### SEMANTIC MODELS OF SIMPLE SENTENCES OF RUSSIAN LANGUAGE AND THEIR APPLICATION IN SYNTACTIC ANALYSIS

G.F. Dyubko, I.S. Karlin

*The main idea of this model is design of analyzer which makes syntactically-semantic analysis of simple sentence of Russian language by determining of sentence semantic functions and by using Λ-dictionary. Semantic function is define sense which is matched to each phrase (group of words that are connected with each other by sense and grammatically) of sentence. Phrase is formed with help of surface sense of words that it contains. This sense is defined by dictionary.*

***Key words:*** *semantics, syntactic, morphology, data base, analysis, case, sentence, natural language, λ-dictionary, semantic function, periphery.*