

ФОРМУЛИРОВКА ЗАДАЧИ СИНТАКСИЧЕСКОГО АНАЛИЗА

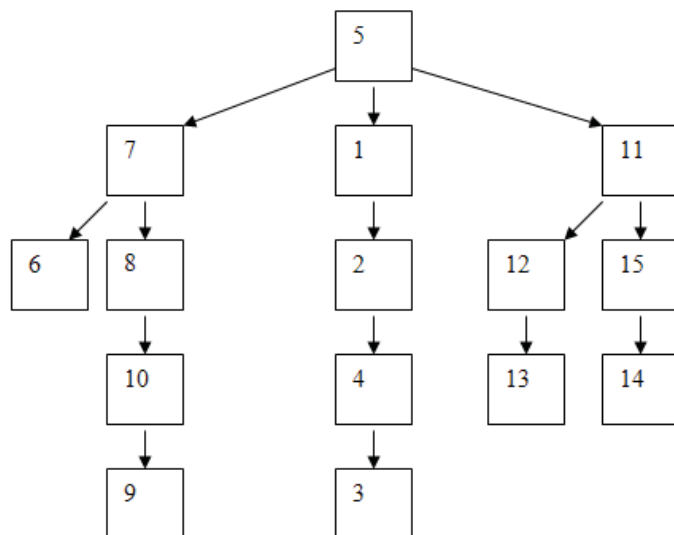
Аннотация: в статье рассматривается решение задачи синтаксического анализа благодаря использованию новых для компьютерной лингвистики подходов к исследованию языка.

Интеллектуальные способности «прививает» компьютеру человек, точнее говоря, та программа, которая разработана и заложена в компьютер человеком. Стремление человека создать программу, наделяющую компьютер языковой способностью, которой обладает он сам, стало в последние полвека движущей силой для исследователей ЕЯ, определило ориентиры и методологию исследования языка, появилась новая область исследований – «компьютерная лингвистика».

Во многих учебных заведениях и научных центрах России немало представителей науки занимаются проблемами компьютерной лингвистики. За короткое время (за два десятилетия) подготовлены различные электронные словари русского языка, экспертные системы, синтаксические анализаторы и программы, переводящие текст из одного языка в другой язык. Однако основная задача компьютерной лингвистики, которая была сформулирована еще А. Тьюрингом – сделать компьютер вполне адекватным собеседником человека – не решена. Компьютер считается адекватным собеседником человека, если последний, общаясь со своим собеседником по сети, не сможет определить по содержанию разговора, с кем он общается: с человеком или компьютером.

Задача в такой постановке становится очень масштабной и сложной. Для ее решения необходимо иметь конструктивную теорию языка, математическую и компьютерную модель языка и, наконец, программу, реализующую эту модель на компьютере.

Неформальная формулировка задачи синтаксического анализа (СА). В качестве первого приближения мы попытаемся сформулировать задачу синтаксического анализа на частном примере. Пусть имеется следующее предложение на русском языке: *соискатель из Дагестанского университета делает интересный доклад по компьютерной лингвистике на факультете управления*. В нашей модели языка предложение по определению представляет собой древовидную структуру. Рассматриваемое предложение имеет следующую структуру (назовем его дерево 1).



Члены предложения пронумерованы слева направо (для экономии записей).

1. соискатель
2. из
3. Дагестанского
4. университета
5. делает
6. интересный

7. доклад
8. по
9. компьютерной
10. лингвистике
11. на
12. факультете
13. управления
14. известного
15. университета

Вершины данного дерева, выписанные в том порядке, в каком они встречаются при его левостороннем обходе, образуют последовательность:

5–7–6–8–10–9– [10–8–7–5]–1–2–4–3–[4–2–1–5]–11–12–13–[12–11]–15–16.

Если удалить повторяющиеся вершины (они указаны в скобках), получится последовательность словоформ (назовем ее ПС1):

делает доклад интересный по лингвистике компьютерной соискатель из университета Дагестанского на факультете управления университета известного.

Процесс построения ПС1 по заданному дереву называется *синтезом предложения*. Другим обходам *дерева* будут соответствовать другие последовательности. Чтобы отличить предложение, представленное деревом, от предложения, заданного последовательностью, вводятся понятия: *предложение–дерево* и *предложение–последовательность*. Одному предложению–дереву может соответствовать несколько предложений–последовательностей. Процесс построения предложения–дерева по заданному предложению–последовательности называется *анализом предложения*.

Программа, анализирующая предложения называется *анализатором*. Предложения выступают как данные, которые обрабатываются анализатором. *Анализатор ЕЯ* должен анализировать любое предложение данного ЕЯ.

Сложность задачи компьютерного анализа заключается в том, что у человека, в частности, у лингвиста и у анализатора – различные требования к точности предмета (данных), с которым они имеют дело. Элементы языка (предложение и его члены), которые определены в грамматике ЕЯ для лингвиста, полны неопределенностей для программиста и для алгоритмизации анализатора.

Прежде чем сформулировать задачу анализа предложения, следует определить само предложение, как структуру, поддающуюся хранению и обработке компьютером. Иначе говоря, необходимо иметь *компьютерную модель предложения*.

Забегая вперед, скажем, что рассматриваемая здесь модель предложения имеет древовидную структуру, между элементами которой должны иметь место отношения двух типов: синтаксические и семантические. Синтаксические и семантические отношения настолько независимы друг от друга, что предложения, удовлетворяющие только синтаксическим отношениям (*синтаксически правильные предложения*) и предложения, удовлетворяющие только семантическим отношениям (*семантически правильные предложения*) могут быть изучены самостоятельно. Более того, они должны быть изучены самостоятельно, ибо, как показывает опыт исследования языка, только так можно определить формальные модели: синтаксически правильного предложения и семантически правильного предложения. Синтаксически и семантически правильное предложение называется *предложением*. В диссертации исследуется *задача синтаксического анализа*, в которой требуется построить синтаксически правильное предложение–дерево, вершинами которого будут словоформы заданной последовательности. Программа, решающая эту задачу, называется *синтаксическим анализатором*.

Алгоритм СА должен дать *точные ответы* на следующие задания:

1. Из данной последовательности словоформ построить словосочетание (которое, в частности, может быть предложением).

2. Из данной последовательности словоформ построить всевозможные предложения.

Чтобы получить точный ответ, вопрос, как минимум, должен быть точно сформулирован. В данном случае необходимо иметь точные определения терминов *словосочетание* и *предложение*. В описании синтаксического анализатора должно быть удалено больше внимания формализации основных понятий синтаксиса ЕЯ и построению математической и компьютерной модели синтаксиса. В таком случае термин программа по отношению к процедуре СА обретает свой истинный смысл. Если ответы синтаксического анализатора на поставленные выше задачи, не соответствуют требованиям грамматики ЕЯ, то источником недоразумения является не анализатор, а модель синтаксиса. Таким образом, *объектом разработки и исследования становится не анализатор, а модель синтаксиса*.

Задачу синтаксического анализа в сформулированном выше виде удастся решать благодаря тому, что были использованы два новых для компьютерной лингвистики подхода к исследованию языка. Во-первых, синтаксиса и семантика были отделены друг от друга. Синтаксиса и семантика представляют собой различные области знания, а язык есть отображение

элементов семантики в элементы синтаксиса. Во-вторых, простое и сложное предложения были отделены друг от друга. Простое и сложное предложения представляют собой различные структуры языка и анализируются различными способами. Работа над диссертацией убедила нас в том, что смешивание этих явлений и конструкций стало роковой ошибкой для компьютерной лингвистики.

Список литературы

1. Грамматика русского языка, т. т. 1 - 3. Издательство Академии Наук СССР, 1960, 720 +702+440 с.
2. Ельмслев Л. Прологомены к теории языка, в кн. Новое в лингвистике, Вып. 1, М., 1960.
3. Мельчук И.А. Опыт теории лингвистических моделей «Смысл – Текст». М., «Наука», 1974, 314 с.
4. Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы) М., 2003, Канд. диссертация.
5. Тестелец Я.Г. Введение в общий синтаксис. М., Российский государственный гуманитарный университет, 2001, 798 с.
6. Тузов В.А. Математическая модель языка. - Л., Изд-во ЛГУ, 1984.