ДРЕВОВИДНЫЕ ГРАММАТИКИ

А. П. Бельтюков, А. Н. Тетерин

Удмуртский государственный универкрситет

Ижевск, Россия

Предлагается альтернатива грамматикам Хомского, покрывающим все возможные классы языков на основе определения языка как дерева объектов, в узлах которого находятся грамматики и атрибуты. Атрибуты определяются функциями, принадлежащими правилам грамматики.

Будем считать, что синтаксические предметы состоят из некоторых *объектов*. Каждый объект может быть построен из других объектов и связей между ними, которые можно считать дополнительными знаками. Эти связи не обязательно выделять как самостоятельные объекты. Эта структура может продолжаться, неограниченно опускаясь вниз в раскрытии объектов, так и поднимаясь вверх и объединяя одни объекты с другими с помощью связей – дополнительных знаков.

В соответствии с традицией, упомянутые дополнительные знаки мы будем называть *терминальными символами*, а названия объектов – *нетерминальными символами*.

Будем придерживаться следующего главного правила таких описаний: нельзя определять объект через самого себя.

Язык определяется деревом объектов, в листьях которого находятся терминальные символы. Это так называемое «дерево описания данных» (ДОД). В процессе анализа входной цепочки ей ставится в соответствие другое дерево объектов – «дерево данных» (ДД). Дерево описания данных – единственно для одного языка и однозначно определяет класс всех возможных деревьев данных («деревьев разбора»), соответствующих предложениям языка.

Объект определяется грамматикой и может иметь атрибуты.

Грамматика – это кортеж $< V_T , V_N , I, P, F >$, где

V_T – некоторое множество, называемое *множеством терминальных символов*,

V_N – некоторое множество, называемое *множеством нетерминальных символов* (названий *дочерних* объектов), множества V_T и V_N не пересекаются,

I – некоторое множество символов, которые называются *символами, вставляемыми во входную цепочку*,

P – множество *редуцирующих правил*, каждое из которых состоит из двух частей: (1) цепочки символов из объединения V_T, V_N и I, (2) цепочки символов из I, длина второй части должна не превосходить длины первой части,

F – отображение, которое ставит в соответствие каждому правилу функцию, областью определения каждой такой функции является некоторое множество кортежей атрибутов, значениями этой функции также являются атрибуты (функции определяют *семантику* этого правила – его *смысл* – обычно смысл связан, с использованием атрибутов дочерних объектов для формирования значений атрибутов объекта). Имя этой функции с её аргументами записывается в дерево разбора в скобочной форме.

Далее определим, что такое *правильный* анализ.

О п р е д е л е н и е 1. Неразрывная часть входной цепочки (подслово) соответствует определённому объекту, если выполнены следующие условия:

1) следующий символ входной цепочки не принадлежит $V_T \cup V_N$ либо его вообще нет, следующий символ входной цепочки не удовлетворяет синтаксису (мы не можем применить правила объекта); выполнено одно из двух:

1-1) применяемые правила *выполнены до конца*,

1-2) применяемые правила выполнены не до конца, *в этом случае формируется символы завершения* (⊥- *конец цепочки*) *для выхода на конец правил; если правила до конца не выполнены*, *необходимо* во входной цепочке вернуться на соответствующее правилам количество символов назад (их количество должно быть меньше на единицу длины части объекта) для исключения этой ситуации (принцип максимизации длины объекта);

2) сформированные на этом уровне атрибуты и атрибуты объектов дочернего уровня удовлетворяют ограничениям этого объекта (если они существуют);

3) следующая часть входной цепочки принадлежит либо терминальным символам вышестоящих уровней, либо объектам того же уровня.

Для любого терминального символа входной цепочки нетрудно установить, что его время разбора (для подходящего естественного вычислителя) будет ограничено некоторой константой (которая будет зависеть от количества объектов, и максимального количества символов в правиле и не будет зависеть от длины входной цепочки). Это означает линейную зависимость времени разбора от длины входной цепочки.

Достоинства нового подхода к синтаксическому анализу:

1) возможность слить лексический, синтаксический и семантический анализ в единый универсальный для каждого уровня дерева объектов (в отличие от обычной схемы количество уровней в дереве не обязательно три) процесс анализа,

2) возможность создать универсальную процедуру поиска и исправления ошибок, в которой каждый вышестоящий уровень корректирует процессы разбора предыдущих уровней.

Синтаксический разбор начинается с применения грамматик, принадлежащих самому нижнему уровню дерева объектов, ведется слева направо, снизу вверх (с элементами анализа сверху вниз) по мере обнаружения объектов нижних уровней делается попытка определить их принадлежность объектам вышележащих уровней иерархии, которые в свою очередь определяют выше лежащие объекты и т.д. При подъеме мы не до конца определяем некоторые объекты в силу недостаточного продвижения анализа по нижнему уровню, но зато получаем список либо терминальных символов либо объектов дочерних уровней, которые нужно распознать в первую очередь.

С точки зрения синтаксического анализа одного уровня все символы в правилах – терминальные. Если рассматривать вопрос о порождении цепочек языка, то он ведется сверху вниз, а объекты в правилах, лежащие на один уровень ниже, рассматриваются как нетерминальные символы.

Возможен синтаксический разбор справа налево. Для параллельной обработки можно сделать следующее:

1) разбить входную цепочку на отрезки, границы которых корректируются во время правостороннего (левостороннего или обоих вместе) разбора,

2) каждый процесс разбора получает свою точку разбора в результате половинного деления или золотого сечения входной цепочки, двигаясь от которой алгоритм последовательно меняет стратегии с левосторонней на правостороннюю и наоборот по мере успешности разбора,

3) при большой глубине дерева объектов входная цепочка разбивается терминальными символами высших уровней (абзацное деление и т. д.).

Рассмотрим пример грамматики объекта «формулы алгебры логики».

Правила (P)	Символ замещения во входной цепочке после применения правила(I)	Функция (F)
1. $\Phi^{A} <= A$	-	-
2. $\Phi^{\neg A} <= \neg A$	A	NOT(A)
3. $\Phi^{K} <= A \wedge A$	A	K(A,A)
4. $\Phi^{D} <= A \vee A$	A	D(A,A)
5. $\Phi^{P} <= (A)$	A	-

$V_T = \{\neg, \wedge, \vee, ()\}$ $V_N = \{A\} I = \{A\}$

Правила для объекта «переменная»:

$\Pi^{имя} <= a$	A	
$\Pi^{имя} <= Aa$	A	Присв(имя)
$\Pi^{конс} <= 0$	C	Присв(знач)
$\Pi^{конс} <= 1$	C	Присв(знач)

V_T = {буква или цифра} $V_N = \{\} I = \{A,C\}$