

В. В. Яковлев, Н. М. Субоч, М. А. Ройтберг, А. Г. Кушниренко  
Пушино, Институт математических проблем биологии РАН, Пушкинский  
госуниверситет, Научно-исследовательский институт системных  
исследований РАН

Проект: Кумир <http://www.infomir.ru>, <http://lpm.org.ru/kumir/>

## Синтаксический разбор программ, содержащих ошибки

### Аннотация

Описание языка программирования определяет поведение компилятора только на правильных программах; поведение компилятора при разборе программ, содержащих синтаксические ошибки (например, диагностика ошибок) обычно никак не регламентировано и оставляется на усмотрение разработчика. В то же время для учебных систем программирования удобная диагностика ошибок имеет первостепенное значение.

Мы предлагаем способ описания поведения компилятора в случае программ, имеющих ошибки в структуре, с помощью подходящей контекстно-свободной грамматики; этот способ был реализован в системе программирования КуМир. Наш опыт показал, что предложенный подход является достаточно гибким и позволяет улучшить качество диагностики.

### Постановка задачи

В настоящей работе мы рассматриваем разбор структуры программ на языке КуМир. КуМир-программа может быть представлена в виде последовательности «канонических» строк, каждая из этих строк — это либо элементарный оператор (присваивание, вызов алгоритма и т. п.), либо компонент операторной скобки (алг, нач, кон, если, то, иначе, все и т. д.).

Под структурой КуМир-программы мы понимаем ее представление в виде последовательности канонических строк, в котором все строки с элементарными операторами считаются эквивалентными (например, заменены специальной строкой простая команда). Структуры правильных КуМир-программ образуют контекстно-свободный язык с несложной грамматикой. В компиляторе системы КуМир анализ структур программы является отдельным модулем.

Нашей целью было разработать язык, который описывает и классифицирует ошибочные структуры программ (по возможности детально) и модуль анализа структуры программы, такие, что:

1. модуль анализа распознает не только правильные программы, но и ошибочные программы в соответствии с разработанным описанием, диагностика ошибок при этом определяется описанием;
2. модуль анализа имеет открытую архитектуру, что позволяет расширять и уточнять классификацию ошибочных ситуаций без вмешательства в исходный код системы КуМир.

Реализация последнего требования позволяет добавление типовых ошибок учеников в систему диагностики непосредственно в ходе учебного процесса.

## Реализация

Грамматики, описывающие как сам язык КуМир  $L(G_0)$ , так и ошибочные ситуации  $L(G_i)$ ,  $i > 0$ , описаны во внешних текстовых файлах в виде набора правил вывода. Данные правила упорядочены по приоритету применения (приоритет указывается в имени файла), поскольку в общем случае множества, порождаемые различными грамматиками  $G_i(\mathcal{H})$  над одним и тем же алфавитом  $\mathcal{H}$ , могут пересекаться. Грамматика, описывающая язык КуМир, то есть все множество правильных программ, имеет наивысший — нулевой приоритет.

С каждым правилом вывода, содержащим терминалы (т. е. символы алфавита  $\mathcal{H}$ , соответствующие типам канонических КуМир-строк) может быть связана некоторая нагрузка, описывающая выполняемые МП-автоматом действия при достижении данного терминального символа. Эти действия представляют собой инструкции на языке ECMAScript, более известном как JavaScript, которые выполняются в случае применения данного правила и могут выдавать сообщения об ошибках, устанавливать ранги отступов, и т. д.

Благодаря такому техническому решению можно легко программировать поведение анализатора в различных ситуациях разбора программ.

## Результаты

В текущем варианте системы КуМир (на момент данной публикации — текущая версия из SVN-репозитория) реализован принципиально новый метод разбора структуры пользовательских программ. Новизна метода заключается в одновременном использовании правил вывода из различных грамматик, порождающих разные языки, в то время как существующие средства генерации анализаторов КС-грамматик, такие как `yacc` или `bison`, могут работать только с одной грамматикой.

Этот метод позволяет, во-первых, корректно диагностировать программы, содержащие более одной ошибки в своей структуре, а во-вторых, распознавать типовые и наиболее распространенные ошибки более крупным планом, так как это сделал бы человек, а не машина.

Открытая система правил вывода ошибочных ситуаций позволяет еще на этапе бета-тестирования «обучить» КуМир-систему распознавать различные ошибочные ситуации, которые допускают учащиеся, для того чтобы впоследствии выдавать соответствующие диагностические сообщения применительно к конкретным ситуациям.

Тем не менее данная работа является экспериментальной, и у разработанного метода нашелся существенный недостаток, а именно длительное время разбора, преодолеть который удалось только ценой ухудшения качества диагностики. Данный метод легко распараллеливается, а тенденция развития современных процессоров ведет к увеличению числа одновременно выполняемых инструкций. Однако экспоненциальная зависимость времени работы данного алгоритма от длины входной цепочки не позволяет надеяться на технический прогресс даже в отдаленной перспективе, поэтому необходимо продолжать исследования в области альтернативных способов разбора программ. Возможным альтернативным подходом является отказ от явного перечисления ошибочных ситуаций и переход к построению правильной программы, наиболее похожей на данную ошибочную.

## Литература

- [1] Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. В двух томах. М.: Мир, 1978. пер. с англ.
- [2] Кушниренко А. Г., Лебедев Г. В., Зайдельман Я. Н. Информатика. 7–9 класс: Учебник для общеобразовательных учреждений. —