

Лабораторная работа №5

«Лексический распознаватель»

Скоробогатов С. Ю., Коновалов А. В.

29 марта 2016

1 Цель работы

Целью данной работы является изучение использования детерминированных конечных автоматов с размеченными заключительными состояниями (лексических распознавателей) для решения задачи лексического анализа.

2 Исходные данные

Пусть лексическая структура модельного языка состоит из четырёх лексических доменов:

1. пробелы — непустые последовательности пробельных символов (пробел, горизонтальная табуляция, маркеры конца строки);
2. идентификаторы — непустые последовательности латинских букв и десятичных цифр, начинающиеся с буквы;
3. целочисленные литералы — непустые последовательности десятичных цифр;
4. ключевые слова (варианты ключевых слов перечислены в таблице 1).
5. знаки операций (варианты знаков операций перечислены в таблице 1).

Чтобы не усложнять лексический анализатор, разрешим идентификаторам примыкать справа к целочисленным литералам.

Таблица 1: Наборы ключевых слов и знаков операций по вариантам

1	if, then, else, (,)	10	def, val, var, [,]
2	do, while, <, >, <>	11	int, float, <, <<
3	begin, end, >, >=	12	fun, let, in, :, ::
4	for, range, ==, !=	13	select, from, :), :(
5	while, wend, <, <=	14	ford, forward, &&,
6	if, fi, do, od, {, }	15	real, longreal, >=, :=
7	case, break, +, ++	16	uint8t, uint128t, (*, *)
8	class, super, -, ->	17	goto, gosub, \(\, \)
9	struct, type, =, ==	18	delete, decltype, ->, <-

3 Задание

Выполнение лабораторной работы состоит из четырёх этапов:

1. построение недетерминированного лексического распознавателя для модельного языка;
2. детерминизация построенного лексического распознавателя и факторизация его алфавита;
3. построение массива обобщённых символов, матрицы переходов и массива заключительных состояний для полученного детерминированного лексического распознавателя с факторизованным алфавитом;
4. разработка лексического анализатора, работающего на основе интерпретации построенных структур данных.

Входной поток для лексического анализатора должен загружаться из файла (в ASCII). В результате работы программы в стандартный поток вывода должны выдаваться описания распознанных лексем в формате

Тег (координаты_фрагмента): изображение_лексемь

При этом лексемь, принадлежащие домену пробелов, должны отбрасываться.

Входной файл может содержать ошибки, при обнаружении которых лексический анализатор должен выдавать сообщение с указанием координаты, восстанавливаться и продолжать работу.

В лабораторной работе разрешается использовать любой язык программирования, поддерживающий массивы с операцией доступа к элементу по индексу, работающей за константное время.