УДК 004.4'412

В.Г. Павлов, канд. тех. наук, доц., К.О. Шапран
Національний технічний університет України "Київський політехнічний інститут", Україна

ОСНОВНІ ПРИНЦИПИ ПОБУДОВИ ЛЕКСИЧНИХ АНАЛІЗАТОРІВ

V. G. Pavlov, Ph.D., Assoc.Prof., K.O. Shapran

BASIC PRINCIPLES OF LEXICAL ANALYZERS GENERATION

Важливим етапом створення компілятора є побудова лексичного аналізатора (сканеру), за допомогою якого у вхідному потоці символів мови здійснюється пошук її окремих структурних одиниць – лексем.

Метою роботи є визначення загальних принципів проектування сканерів.

Лексичний аналіз є першою фазою роботи компілятора, під час якої здійснюється підготовка для виконання аналізу синтаксису тексту. Для цього визначаються типи усіх знайдених лексем, які відповідають певним структурам мови – токенам. Таким чином, сканер ніби «огороджує» синтаксичний аналізатор від безпосередньої роботи з лексемами, чим значно спрощує його роботу і підвищує швидкодію та ефективність компілятора в цілому [1].

Тож, можна сформулювати основні принципи роботи лексичних аналізаторів:

1. Зчитування потоку символів з тексту вхідної програми.
2. Видалення з тексту «зайвих» символів, які не входять до складу токенів – пробіл, символи табуляції та нового рядка, тощо.
3. Групування отриманих символів у лексеми, що включає визначення границь кожної лексеми.
4. Виявлення та повідомлення про помилку, якщо лексема неправильна.
5. Формування і заповнення таблиць лексем та ідентифікаторів для їх наступної передачі синтаксичному аналізатору [2].

Зазначимо основні відмінності між таблицею лексем та ідентифікаторів, а саме:

1) Таблиця лексем включає усі їх можливі типи. Таблиця ідентифікаторів містить тільки визначені типи лексем – ідентифікатори та константи.

2) Будь-яка лексема, в таблиці лексем, може зустрічатися необмежену кількість разів. У таблиці ідентифікаторів кожна лексема (ідентифікатор чи константа) визначена тільки один раз.

3) Обов'язково в таблиці лексем вони розміщуються в тому ж порядку, що й у вхідній програмі, а в таблиці ідентифікаторів – розташовані в порядку, який забезпечує їх зручний пошук.

Розглянемо процес лексичного аналізу та формування таблиць на прикладі.

Приклад. При лексичному аналізі фрагменту вхідного коду на мові Java:

if (flag) return 1;

– отримані наступні таблиці: таблиця лексем, без урахування пробільних символів (таблиця 4), таблиця ідентифікаторів (таблиця 3) і таблиця літералів (таблиця 2).

Також наведена частина таблиці термінальних символів мови програмування Java (таблиця 1), яка застосовувалася під час лексичного аналізу наведеного виразу, у ній зберігається інформація про основні конструкції мови (ключові слова, розділювачі, знаки операцій і т.п.).

У заданому виразі до термінальних символів відносяться: if, (,), return, ; та пробіл.

Таблиця 1 – Таблиця термінальних символів

№ символа	Символ	Тип
1	;	Розділювач
2	|…|	Розділювач
3	(	Розділювач
4	)	Розділювач
5	if	Ключове слово
6	return	Ключове слово

Таблиця констант включає записи наступної структури: номер константи в таблиці констант; рядкове представлення константи; тип (цілий, дійсний, логічний, символьний, рядковий); точність представлення константи в пам'яті (у вигляді числа байтів, відведених для зберігання).

До констант вхідного виразу запишемо лексему 1 (тип int).

Таблиця 2 – Таблиця констант

№	Ім'я	Тип	Значення
1	«1»	int	00000001

Ідентифікатори, які складають відповідну таблицю, мають атрибути: номер, ім'я (змінної, масиву, функції й т.п.), тип (дійсний, цілий, логічний, рядковий, символьний) та адреса в пам'яті.

Виділимо ідентифікатори заданої послідовності символів – flag (тип boolean).

Таблиця 3 – Таблиця ідентифікаторів змінних

№	Вид (ім'я)	Тип	Адреса
1	flag	boolean	∅ (поки що не розподілена)

Вихідною таблицею лексичного аналізатора є таблиця лексем, що одночасно є вхідною для синтаксичного аналізу. Код лексеми складається з ознаки типу лексеми та номеру лексеми в таблиці лексем. Позначення типів лексем: I – для лексем-ідентифікаторів, C – для лексем-констант, T – для лексем-термінальних символів мови.

Таблиця 4 – Таблиця лексем

№	Лексема	Тип лексеми	Код	Посилання
1	if	Службове слово	T5	Адреса в таблиці терміналів
3	(	Службовий символ	T3	Адреса в таблиці терміналів
4	flag	Ідентифікатор	I1	Адреса в таблиці ідентифікаторів
5	)	Службовий символ	T4	Адреса в таблиці терміналів
6	return	Службовий символ	T6	Адреса в таблиці ідентифікаторів
7	1	Літерал	C1	Адреса в таблиці літералів
8	;	Службовий символ	T1	Адреса в таблиці терміналів

Отже, ми визначили та висвітлили основні принципи побудови лексичних аналізаторів. Виокремили особливості деяких етапів роботи сканера. Обґрунтували зв'язок лексичного аналізатора з іншими частинами компілятора.

Література

1. Компиляторы: принципы, технологии и инструментарий / [Альфред В. Ахо, Моника С. Лам, Рави Сети, Джеффри Д. Ульман] ; [Пер. с англ.]. – 2-е изд. – М.: ООО «И. Д. Вильямс», 2008. – 155-158 с.

2. Системное программное обеспечение: [учебник для вузов] / Молчанов А. Ю.. – 3-е изд. – СПб.: Питер, 2010. – 88-89 с.