**Алфавит**

Определение класса лексем языка Си

**Пример класса «идентификаторов» языка Си**

Идентификатор – произвольное слово вида *px*, где *p* – односимвольное слово в алфавите из букв, *x* – произвольное слово, составленное из букв и цифр. Под буквой подразумевается любая прописная или строчная буква , а также знак «подчеркивания» расширенной ASCII кодовой страницы .

[\_A-Za-z] – алфавит, в котором представимы слова вида p. Поскольку символы отождествляются с односимвольными словами, тогда алфавит представляет все слова вида р.

[\_A-Za-z0-9] – алфавит, в котором представимы слова вида *x*.

[\_A-Za-z0-9]\* -- результат применения оператора Клини к алфавиту определяет все слова в данном алфавите ( слова произвольной длины, включая «пустое» слово).

[\_A-Za-z] [\_A-Za-z0-9]\* -- сцепление (конкатенация) двух языков дает искомый язык: класс идентификаторов языка Си.

[\_A-Za-z0-9] – допустимый алфавит, в котором определяются идентификаторы.

[^\_A-Za-z0-9] – не допустимый алфавит класса идентификаторов относительно входного алфавита (ASCII).

**Пример класса «целых восьмеричных констант» языка Си**

Константа – произвольное слово вида *pxs*, где *p* – префикс, фиксированное слово, состоящей из цифры «нуль»; *x* – произвольное слово, составленное из восьмеричных цифр; *s* – суффикс. Под суффиксом понимается либо пустое слово, либо слово, составленное из одной из букв *l* или *u*, либо двухбуквенные слова *lu* или *ul*.   
Вместо строчных букв можно использовать прописные.

[0] – алфавит, в котором представимы слова вида p.

[0-7] – алфавит, в котором представимы слова вида *x*.

[0-7]\* -- все слова вида *x*

[lLuU] – алфавит и определение всех односимвольных суффиксов.

[Ll][Uu] | [Uu][Ll] – определение всех двухсимвольных суффиксов.

[lLuU] | [Ll][Uu] | [Uu][Ll] – определение всех не пустых суффиксов.

([lLuU] | [Ll][Uu] | [Uu][Ll])? – определение суффиксов.

[0] [0-7]\* ([lLuU] | [Ll][Uu] | [Uu][Ll])? -- -- сцепление (конкатенация) трех языков дает искомый язык: класс **«целых восьмеричных констант» языка Си.**

[lLuU0-7] – допустимый алфавит.

[^lLuU0-7] – не допустимый алфавит.

**Разбиение алфавита**

Символы *a* и *b* эквивалентны относительно языка *L*, если для любых слов *x* и *y*

слово *xay* принадлежит языку *L* тогда и только тогда, когда *xby* принадлежит языку *L*.

Записываем *a* ≅ *b* или *a* = *b* (mod L), если символы *a* и *b* эквивалентны относительно языка *L*.

**Пример класса «идентификаторов» языка Си**

[^\_A-Za-z0-9] -- не допустимые символы определяют отдельный класс эквивалентных

символов;

[\_A-Za-z0-9] – множество допустимых символов образуют два класса эквивалентных

символов: [\_A-Za-z] и [0-9].

Все символы входного алфавита (ASCII) относительно языка -- **класса «идентификаторов» языка Си** распадаются на три класса эквивалентных символов:

[\_A-Za-z] , [0-9] , [^\_A-Za-z0-9]. Причем символы из разных классов не эквивалентны.

**Упражнение 1.** Провести рассуждение (привести контрпримеры, доказывающие не эквивалентность символов, принадлежащих разным классам).

**Упражнение 2.** Провести рассуждение, пользуясь правилом подстановки, для доказательства эквивалентности символов, принадлежащих одному классу.

**Класс «идентификаторов» языка Си** определяется в нормальной форме, основанной на **разбиении алфавита:**

[\_A-Za-z] ( [\_A-Za-z] | [0-9])\*

Разбиение входного алфавита:

[\_A-Za-z] , [0-9] , [^\_A-Za-z0-9].

Выбор абстрактного алфавита из представителей классов эквивалентных символов:

a=[\_A-Za-z];

9=[0-9];

?= [^\_A-Za-z0-9];

Абстрактный алфавит: B = {a, 9, ?}

Представление **класса «идентификаторов» языка Си** в алфавите B:

[a] [a9]\* или проще – a(a|9)\*

**Класс «целых восьмеричных констант» языка Си**

определяется в нормальной форме, основанной на **разбиении алфавита**:

[0] ( [0] | [1-7]\* ([Ll] | [Uu] | [Ll][Uu] | [Uu][Ll])?

Разбиение входного алфавита:

[0], [1-7], [Uu], [Ll], [^LlUu0-7]

Выбор абстрактного алфавита из представителей классов эквивалентных символов:

0=[0];

7=[1-7];

u=[Uu];  
l=[Ll];

?= [^LlUu0-7];

Абстрактный алфавит: B = {0, 7, u, l, ?}

Представление **класса «целых восьмеричных констант» языка Си** в алфавите B:

[0] [07]\*([ul] | ul | lu)? или проще -- 0(0|7)\*(u|l | ul | lu)?