# Compiladores I

Definición del Lenguaje `NanoPascal`

Iván de Jesús Deras Tábora

Universidad Tecnlógica Centroamericana (UNITEC)

05/26/2018

# 1 Introducción

El proyecto de la clase consite en implementar un intérprete para un lenguaje llamado `NanoPascal`. El cual es un lenguaje fuertemente tipificado, con soporte para tipos de datos enteros, booleanos, caracter y arreglos unidimensionales. `NanoPascal` es un subconjunto del lenguaje Pascal. Debemos tener en cuenta que `NanoPascal` no es una copia exacta del lenguaje Pascal. La funcionalidad de `NanoPascal` ha sido reducida considerablemente comparado con un lenguaje "completo". Esto se hizo para hacer el proyecto de la clase realizable en un trimestre (10 semanas). A pesar de esto `NanoPascal` será capaz de ejecutar programas complicados como el siguiente:

# 2 Programa de ejemplo en NanoPascal

NanoPascal Sample Code

```
program GCD;
var
    a, b: Integer;
    x, y, z: Integer;

// Funcion que calcula el maximo comun divisor
function gcd(a: Integer; b: Integer): Integer;
begin
    if b = 0 then
        gcd := a
    else
        gcd := gcd(b, a mod b);
end;

begin
    a := 10;
    b := 20;
    x := a;
    y := b;
    z := gcd(x, y);

    writeln(z);
end.
```

# 3 Notación a utilizar

| | |
|---|---|
| `<nts>` | Significa que ***nts*** es un símbolo no terminal. |
| `ts` | Significa que ts es un símbolo terminal, o sea un Token reconocido por el analizador léxico. |
| `'x'` | Significa que la cadena x es un terminal cuyo lexema es x. |
| `[r]` | Significa cero o una ocurrencia de $r$. |
| `r*` | Significa cero o más ocurrencias de $r$. |
| `r+` | Significa una o más ocurrencias de $r$. |
| `{r}+,` | Una lista separada por comas de una o más ocurrencias de $r$. |
| `{}` | Las llaves son usadas para agrupar |
| `\|` | Usado para separar alternativas |
| `[c1-c2]` | Denota un conjunto de caracteres. Ej: **[a-c]** denota los tres caracteres **a, b, c** |

# 4 Consideraciones Léxicas

Las palabras reservadas y los identificadores son case-insensitive. Por ejemplo, `if` es una palabra reservada, pero también lo son `If`, `IF` y todas sus variantes. Además, `compi2` y `Compi2` hacen referencia al mismo identificador. Hay dos tipos de comentarios:

- De linea: Comienzan con `//` y terminan con un fin de línea.
- De bloque:
  - Comienzan con `(*` y terminan con `*)`
  - Comienzan con `{` y terminan con `}`

`NanoPascal` soportará las siguientes directivas de compilación condicional:

- `{$ifdef X}` para verificar si el símbolo `X` ha sido definido.
- `{$ifndef X}` para verificar si el símbolo `X` no ha sido definido.

Ambas directivas soportarán el uso de `{$else}` y terminarán con la directiva `{$endif}`. Por defecto `NanoPascal` definirá el símbolo `NANOPASCAL`. La siguiente figura muestra un ejemplo del uso de estas directivas:

NanoPascal Directives

```
program Directives;
begin
{$ifdef NANOPASCAL}
  writeln('Hello World from NanoPascal');
{$else}
  writeln('Hello World from Other Pascal Compiler');
{$endif}
end.
```

Si ejecutamos este programa utilizando `NanoPascal` la salida debería ser: `Hello World from NanoPascal`, pero si lo ejecutamos utilizando otro compilador la salida debería ser: `Hello World from Other Pascal Compiler`.

# 5 Definición de Tokens

Las palabras reservadas son las siguientes:

```
program var array of integer boolean char not and or xor shl shr div mod begin end break if then
else while repeat until for do write writeln
```

Los operadores y tokens de puntuación son los siguientes:

```
[ ] , ; ( ) = - + * < > <> <= >=
```

Tokens como `stringConstant` el cual define una constante cadena como `'hello, world'` no aparecen en la lista anterior pero son tokens válidos y son usados en la definición de la gramática de `NanoPascal` (Vea la sección 5).

Los identificadores, denotados por el token `ID` son definidos como una cadena de caracteres que comienza con un carácter alfabético o un guión bajo, seguido de cero o más caracteres alfanuméricos incluyendo el guión bajo.

Las palabras reservadas y los identificadores deben estar separados por espacios en blanco, o un token que no sea una palabra reservada o un identificador. Ej: `varwhiletrue` es un solo identificador, no tres distintas palabras reservadas. Veas la sección 4.2 para algunos ejemplos.

Constantes de cadena, denotadas por el token `stringConstant` tendrán un valor léxico compuesto de caracteres encerrados entre comillas sencillas. Una cadena debe comenzar y terminar en una sola línea, no puede expandirse en multiples lineas. Para más detalles sobre cadenas y caracteres vea la sección 4.3

Contantes numéricas en `NanoPascal`, denotadas por el token `intConstant`, son decimales (base 10), hexadecimales (base 16) o binarias (base 2). Una constante entera hexadecimal comienza con `$` seguido de una secuencia de dígitos hexadecimales. Las constantes binarias comienzan con el símbolo `%` . Ejemplos de constantes enteras:

```
8, 012, $0, $12aE, %01100101
```

Constantes Carácter, denotados por el token `charConstant` tendrán un valor léxico que es un solo character encerrado entre comillas sencillas. Una constante carácter es cualquier caraćter ASCII que sea imprimible (Valores ASCII enter 32 y 126). Una constante carácter no puede ser una comilla sencilla `'''`. Para más detalles sobre cadenas y caracteres vea la sección 4.3.

# 6 Reconocimiento de Tokens y Espacios en Blanco

El reconocimiento de tokens tales como constants enteras, palabras reservadas e identificadores son explicadas usando las siguientes reglas. En efecto estas reglas definen un algoritmo para agrupar caracteres del conjunto [a-zA-Z0-9] en tokens.

- Si la secuencia comienza con `$`, entonces este caracter y la secuencia más larga de caracteres del conjunto `[0-9a-fA-F]` que sigue forman una constante hexadecimal. El último carácter de esa secuencia marca el fin del token.
- Si la secuencia comienza con `%`, entonces este caracter y la secuencia más larga de caracteres del conjunto `{0, 1}` que sigue forman una constante binaria. El último carácter de esa secuencia marca el fin del token.
- Si la secuencia comienza con un dígito decimal entonces la secuencia más larga de dígitos decimales forman una constante entera. Tenga en cuenta que la semántica de verificación de rango se llevará a cabo luego, de esta forma la secuencia 123456789123456789 la cual claramente está fuera de rango, será reconocida como un solo token por el lexer.
- Si la secuencia comienza con un carácter alfabético ó `_`, entonces este caracter y la secuencia más larga de caracteres alfanuméricos `[0-9a-zA-Z_]` que siguen a este caracter inicial forman un token que puede ser un identificador o una palabra reservada.
- Espacios en blanco y otras definiciones de tokens juegan un papel importante en la delimitación de tokens. Por ejemplo la cadena `shl3` es un solo identificador, pero `shl 3` son dos tokens, la palabra reservada `shl` y el token constante entera 3, `shl(3)` representa 4 tokens, la palabra reservada `shl`, el paréntesis izquierdo, la constante entera `3` , y el paréntesis derecho. Se consideran espacios en blanco los caracteres de fin de linea, tabulador y el carácter ASCII de espacio en blanco.

Aquí hay varios ejemplos de las reglas explicadas anteriormente:

```
$123food = INT_CONST(123f, 16), IDENTIFIER(ood)
$food123 = INT_CONST(f, 16), IDENTIFICADOR(ood123)
123break = INT_CONST(123, 10), KEYWORD(BREAK)
$123shl3 = INT_CONST(123, 16), IDENTIFICADOR(rot3)
$123shl 3 = INT_CONST(123, 16), KEYWORD(SHL), INT_CONST(3, 10)
1250x356 = INT_CONST(1250, 10), IDENTIFICADOR(x356)
break123 = IDENTIFICADOR(break123)
breakwhile = IDENTIFICADOR(breakwhile)
```

# 7 Constantes de cadena y carácter

Las constantes de cadenas, denotadas por el token `stringConstant` tendrán un valor léxico compuesto de los caracteres encerrados entre comillas sencillas. Una cadena debe comenzar y terminar en una sola línea, no puede dividirse en múltiples líneas. Para poder insertar una comilla sencilla en una literal de cadena utilizaremos doble comilla sencilla. Las constantes de carácter, denotadas por el token `charConstant` tendrán un valor léxico que es un solo carácter encerrado entre comillas sencillas. Una constante carácter es cualquier símbolo ASCII que tenga representación gráfica (Valores entre 32 y 126). Una constante carácter no puede ser una comilla sencilla ’’’ . Las constantes de carácter pueden utilizar la secuencia de escape doble comilla sencilla para definir una comilla sencilla como literal caracter. Al momento de reconocer constantes de carácter `charConstant` o constantes de cadena `stringConstant` en el analizador léxico deberá tener en cuenta lo siguiente:

- Las constantes de cadena y caracter pueden contener la secuencia de escape doble comilla
- Las constante de cadena y de carácter deberán cerrarse con una comilla sencilla, en particular las siguientes constantes deberán tratarse como un error:
  - Constantes de caracter que contengan cero caracteres ’’
  - Constantes de cadena y de carácter sin el carácter de cierre deben ser reportadas como errores.

# 8 Gramática de Referencia

La siguiente gramática de referencia define la estructura de un programa de `NanoPascal`. Aquí usamos la notación definida en la sección 3. Esta gramática de referencia no es libre de contexto, aunque podría convertirse fácilmente en gramática libre de contexto.

```
   <program> -> program ID ';' [variable-section] <subprogram-decl>* begin [ {<statement>}+; ] [';']
end '.'
   <variable-section> -> var <variable-decl>+
   <variable-decl> -> {id}+, ':'  <type> ';'
   <type> -> integer | boolean | char | <array-type>
   <array-type> -> array '[' intConst '..'  intConst ']' of <type>
   <subprogram-decl> -> <subprogram-header> [variable-section] begin <statement>* end ';'
   <subprogram-header> -> <function-header> | <procedure-header>
   <function-header> -> function ID [ '(' [ <argument-decl>}+; ] ')' ] ':'  <type> ';'
   <function-header> -> procedure ID [ '(' [ <argument-decl>}+; ] ')' ] ';'
   <argument-decl> -> var ID ':' <type> | ID ':' <type>
   <statement> -> <assign>
                | <subprogram-call>
                | if <expr> then <block> [else <block>]
                | while <expr> do <block>
                | repeat <block> until <expr>
                | for <assign> <expr> to <expr> do <block>
                | break
                | continue
   <block> -> <statement>
             | begin [ {<statement>}+; ] [';'] end
   <assign> -> <lvalue> ':=' <expr>
   <subprogram-call> -> ID [ '(' [ {<expr>}+, ] ')' ]
                      | write '(' {<argument>}+, ')'
                      | writeln [ '(' [ {<argument>}+, ] ')' ]
                      | read '(' {<argument>}+, ')'
   <argument> -> stringConstant | <expr>
   <lvalue> -> ID | ID '[' <expr> ']'
   <expr> -> <lvalue>
            | <subprogram-call>
            | <constant>
            | <expr> <bin-op> <expr>
            | '-' <expr>
            | not <expr>
            | '(' <expr> ')'
   <bin-op> -> <arith-op> | <rel-op> | <eq-op> | <cond-op>
   <arith-op> -> '+' | '-' | '*' | div | mod | shl | shr | '<<' | '>>'
   <rel-op> -> '<' | '>' | '<=' | '>='
   <eq-op> -> '=' | '<>'
   <cond-op> -> and | or | xor
   <constant> -> intConstant | charConstant | <bool-constant>
   <bool-constant> -> true | false
```

# 9 Reglas Semanticas

Un programa de `NanoPascal` consiste de una definición de un programa asociado con un identificador. La declaración del programa consiste de declaraciones de variableas y declaraciones de subprogramas a nivel global. Las variables globales pueden ser accesadas globalmente por todos los subprogramas.

## 9.1 Tipos de Datos

Hay tres tipos básicos en NanoPascal – Integer para enteros, Boolean para booleanos y Char para caracteres. Además el lenguage soporta arreglos unidimensionales de estos tipos básicos. Todos los arreglos tienen un tamaño fijo definido en tiempo de compilación.

## 9.2 Expresiones

Las expresiones siguen las reglas de Pascal para el orden de evaluación. Las constantes enteras evaluan a su valor entero. Las contantes carácter evaluan a su valor entero ASCII, ej. 'A' evalua a 65 (si tiene Linux puede ejecutar man ascii para la tabla ASCII completa). Una expresión que hace referencia a un elemento de un arreglo, ej. x[10] evalua al valor contenido en la posición referenciada. Los operadores relacionales son usados para comparar expresiones de tipo entero, caracter y booleano. Los operadores de igualdad '=' y '<>' están definidos para los tipos Integer, Char, y Boolean y pueden ser utilizados para comparar dos expresiones que tengan el mismo tipo. El resultado de un operador relacional o de igualdad tiene tipo Boolean. Los operadores booleanos and y or son interpretados usando corto-circuito asi como en Pascal. Esto significa que los efectos secundarios del segundo operando no son ejecutados si el resultado del primer operando determina el valor de la expresión (Esto es: si el resultado es falso para and o verdadero para or).

| Operadores | Precedencia |
|---|---|
| not, unary - | La más alta |
| * div mod and shl shr << >> | Segundo nivel |
| + - or xor | Tercer nivel |
| = <> < > <= >= | Nivel más bajo |

**Precedencia de operadores en NanoPascal**

El nível de precedencia de cada operador se muestra en la tabla anterior. Todos los operadores en el mismo nivel tienen la misma precedencia. Los operadores con igual precedencia se asocian por la izquierda.

## 9.3 Sentencias

Las expressiones condicionales en sentencias como If, While y Repeat deben evaluar al tipo Boolean, en caso de no ser asi el compilador deberá generar un error.

Las asignaciones solo son válidas para tipos de datos Integer, Char, y Boolean, esto implica que no se pueden asignar arreglos.