# Contenido

# 4.5 RECUPERACIÓN DE ERRORES EN ANALIZADORES SINTÁCTICOS DESCENDENTES

La respuesta de un analizador sintáctico a los errores de sintaxis a menudo es un factor crítico en la utilidad de un compilador. Un analizador sintáctico debe determinar por lo menos si un programa es sintácticamente correcto o no. Un analizador sintáctico que sólo realiza esta tarea se denomina **reconocedor**, puesto que se limita a reconocer cadenas en el lenguaje libre de contexto generado por la gramática del lenguaje de programación en cuestión. Vale la pena establecer que cualquier analizador sintáctico debe comportarse por lo menos como un reconocedor; es decir, si un programa contiene un error de sintaxis, el analizador sintáctico debe indicar que existe *algún* error y, a la inversa, si un programa no contiene errores de sintaxis, entonces el analizador sintáctico no debe afirmar que existe alguno.

Más allá de este comportamiento mínimo, un analizador sintáctico puede mostrar muchos niveles diferentes de respuestas a los errores. Habitualmente, un analizador sintáctico intentará proporcionar un mensaje de error importante, cuando menos para el primer error que encuentre, y también intentar determinar de manera tan exacta como sea posible la ubicación en la que haya ocurrido el error. Algunos analizadores sintácticos pueden ir tan lejos como para intentar alguna forma de **corrección de errores** (o, para decirlo de manera quizás más apropiada, **reparación de errores**), donde el analizador sintáctico intenta deducir un programa correcto a partir del incorrecto que se le proporciona. Si intenta esto, la mayor parte de las veces se limitará sólo a casos fáciles, como la falta de un signo de puntuación. Existe un grupo de algoritmos que se pueden aplicar para encontrar un programa correcto que en cierto sentido se parezca mucho al proporcionado (habitualmente en términos del número de tokens que deben insertarse, eliminarse o modificarse). Esta **corrección de errores de distancia mínima** es por lo regular muy ineficiente como para aplicarse a cualquier error y, en cualquier caso, la reparación de errores que da como resultado a menudo está muy lejos de lo que el programador pretendía. Por consiguiente, es raro que se vea en analizadores sintácticos reales. A los escritores de compiladores les es muy difícil generar mensajes de error significativos sin intentar hacer corrección de errores.

La mayoría de las técnicas para la recuperación de errores son de propósito específico, ya que se aplican a lenguajes específicos y a algoritmos de análisis sintáctico específico, con muchos casos especiales para situaciones particulares. Los principios generales son difíciles de obtener. Algunas consideraciones importantes que se aplican son las siguientes.

1. Un analizador sintáctico debería intentar determinar que ha ocurrido un error *tan pronto como fuera posible*. Esperar demasiado tiempo antes de la declaración del error significa que la ubicación del error real puede haberse perdido.
2. Después de que se ha presentado un error, el analizador sintáctico debe seleccionar un lugar probable para reanudar el análisis. Un analizador sintáctico siempre debería intentar analizar tanto código como fuera posible, a fin de encontrar tantos errores reales como sea posible durante una traducción simple.
3. Un analizador sintáctico debería intentar evitar el **problema de cascada de errores**, en la cual un error genera una larga secuencia de mensajes de error falsos.
4. Un analizador sintáctico debe evitar bucles infinitos en los errores, en los que se genera una cascada sin fin de mensajes de error sin consumir ninguna entrada.

Algunos de estos objetivos entran en conflicto entre sí, de tal manera que un escritor de compiladores tiene que efectuar "convenios" durante la construcción de un manejador de errores. Por ejemplo, el evitar los problemas de cascada de errores y bucle infinito puede ocasionar que el analizador sintáctico omita algo de la entrada, con lo que compromete el objetivo de procesar tanta información de la entrada como sea posible.

## 4.5.1 Recuperación de errores en analizadores sintácticos descendentes recursivos

Una forma estándar de recuperación de errores en los analizadores sintácticos descendentes recursivos se denomina **modo de alarma**. El nombre se deriva del hecho que, en situaciones

complejas, el manejador de errores consumirá un número posiblemente grande de tokens en un intento de hallar un lugar para reanudar el análisis sintáctico (en el peor de los casos, incluso puede consumir todo el resto del programa, lo que no es mejor que simplemente salir después del error). Sin embargo, cuando se implementa con cuidado, éste puede ser un método para la recuperación de errores mucho mejor que lo que implica su nombre.[4] Este modo de alarma tiene, además, la ventaja de que virtualmente asegura que el analizador sintáctico no caiga en un bucle infinito durante la recuperación de errores.

El mecanismo básico del modo de alarma es proporcionar a cada procedimiento recursivo un parámetro extra compuesto de un conjunto de **tokens de sincronización**. A medida que se efectúa el análisis sintáctico, los tokens que pueden funcionar como tokens de sincronización se agregan a este conjunto conforme se presenta cada llamada. Si se encuentra un error, el analizador sintáctico **explora hacia delante**, desechando los tokens hasta que ve en la entrada uno del conjunto de tokens de sincronización, en donde se reanuda el análisis sintáctico. Las cascadas de errores se evitan (hasta cierto punto) al no generar nuevos mensajes de error mientras tiene lugar esta exploración adelantada.

Las decisiones importantes que se tienen que tomar en este método de recuperación de errores consisten en determinar qué tokens agregar al conjunto de sincronización en cada punto del análisis sintáctico. Por lo general los conjuntos Siguiente son candidatos importantes para tales tokens de sincronización. Los conjuntos Primero también se pueden utilizar para evitar que el manejador de errores omita tokens importantes que inicien nuevas construcciones principales (como sentencias o expresiones). Los conjuntos Primero también son importantes, puesto que permiten que un analizador sintáctico descendente recursivo detecte pronto los errores en el análisis sintáctico, lo que siempre es útil en cualquier recuperación de errores. Es importante darse cuenta que el modo de alarma funciona mejor cuando el compilador "sabe" cuándo *no* alarmarse. Por ejemplo, los símbolos de puntuación perdidos, tales como los de punto y coma, y las comas, e incluso los paréntesis derechos olvidados, no siempre deberían provocar que un manejador de errores consuma tokens. Naturalmente, debe tenerse cuidado para asegurar que no se presente un bucle infinito.

Ilustraremos la recuperación de errores en modo de alarma esquematizando en pseudocódigo su implementación en la calculadora descendente recursiva de la sección 4.1.2 (véase también la figura 4.1). Además de los procedimientos *match* y *error*, que en esencia permanecen iguales (excepto que *error* ya no sale de inmediato), tenemos dos procedimientos más, *checkinput*, que realiza la verificación temprana de búsqueda hacia delante, y *scanto*, que es el token consumidor en modo de alarma propiamente dicho:

```
procedure scanto ( synchset ) ;
begin
  while not ( token in synchset ∪ { $ }) do
    getToken ;
end scanto ;

procedure checkinput ( firstset, followset ) ;
begin
  if not ( token in firstset ) then
    error ;
    scanto ( firstset ∪ followset ) ;
  end if ;
end;
```

Aquí el signo $ se refiere al fin de la entrada (EOF).

4. Wirth [1976], de hecho, llama al modo de alarma la regla de "no alarma", supuestamente en un intento de mejorar su imagen.

Estos procedimientos se utilizan como sigue en los procedimientos *exp* y *factor* (que ahora tienen un parámetro *synchset*):

```
procedure exp ( synchset ) ;
begin
  checkinput ( { (, number }, synchset ) ;
  if not ( token in synchset ) then
    term ( synchset ) ;
    while token = + or token = - do
      match (token) ;
      term ( synchset ) ;
    end while ;
    checkinput ( synchset, { (, number }) ;
 end if;
end exp ;

procedure factor ( synchset ) ;
begin
 checkinput ( { (, number }, synchset ) ;
 if not ( token in synchset ) then
  case token of
  ( : match( ( ) ;
      exp ( { ) } ) ;
      match( ) ) ;
  number :
      match(number) ;
  else error ;
  end case ;
  checkinput ( synchset, { (, number }) ;
 end if ;
end factor ;
```

Advierta cómo *checkinput* es llamado dos veces en cada procedimiento: una vez para verificar que un token en el conjunto Primero sea el token siguiente en la entrada y una segunda vez para verificar que un token en el conjunto Siguiente (o *synchset*) sea el token siguiente en la salida.

Esta forma de modo de alarma generará errores razonables (mensajes de error útiles que también se pueden agregar como parámetros a *checkinput* y *error*). Por ejemplo, la cadena de entrada `(2+-3)*4-+5` generará exactamente dos mensajes de error (uno para el primer signo de menos y otro para el segundo signo de más).

Observamos que, en general, *synchset* se pasa en las llamadas recursivas, con nuevos tokens de sincronización agregados de manera apropiada. En el caso de *factor*, se hace una excepción después de que se ve un paréntesis izquierdo: se llama a *exp* con paréntesis derecho sólo como su conjunto siguiente (*synchset* se descarta). Esto es típico de la clase de análisis de propósito específico que acompaña a la recuperación de errores en modo de alarma. (Hicimos esto de modo que, por ejemplo, la expresión `(2+*)` no generase un falso mensaje de error para el paréntesis derecho.) Dejamos un análisis del comportamiento de este código, así como su implementación en C, para los ejercicios. Por desgracia, para obtener los mejores mensajes de error y recuperación de errores, virtualmente toda prueba de token se debe examinar por la posibilidad de que una prueba más general, o una prueba más temprana, mejore el comportamiento del error.

### 4.5.3 Recuperación de errores en el analizador sintáctico de TINY

El manejo de errores del analizador sintáctico de TINY, como se da en el apéndice B, es muy rudimentario: sólo está implementada una forma muy primitiva de recuperación en modo de alarma, sin los conjuntos de sincronización. El procedimiento `match` simplemente declara error, estableciendo cuál token que no esperaba encontró. Además, los procedimientos `statement` y `factor` declaran error cuando no se encuentra una selección correcta. El procedimiento `parse` también declara error si se encuentra un token distinto al fin de archivo después de que termina el análisis sintáctico. El principal mensaje de error que se genera

es "unexpected token" ("token inesperado"), el cual puede no ser muy útil para el usuario. Además, el analizador sintáctico no hace un intento por evitar cascadas de error. Por ejemplo, el programa de muestra con un signo de punto y coma agregado después de la sentencia write

```
...
 5: read x ;
 6: if 0 < x then
 7:    fact := 1;
 8:    repeat
 9:       fact := fact * x;
10:       x := x - 1
11:    until x = 0;
12:    write fact; {<-- ¡SIGNO DE PUNTO Y COMA ERRÓNEO! }
13: end
14:
```

provoca que se generen los siguientes *dos* mensajes de error (cuando únicamente ha ocurrido un error):

```
Syntax error at line 13: unexpected token -> reserved word: end
Syntax error at line 14: unexpected token -> EOF
```

Y el mismo programa con la comparación < eliminada en la segunda línea de código

```
...
 5: read x ;
 6: if 0 x then { <-- ¡SIGNO DE COMPARACIÓN PERDIDO AQUÍ! }
 7:    fact := 1;
 8:    repeat
 9:       fact := fact * x;
10:       x := x - 1
11:    until x = 0;
12:    write fact
13: end
14:
```

ocasiona que se impriman *cuatro* mensajes de error en el listado:

```
Syntax error at line 6: unexpected token -> ID, name = x
Syntax error at line 6: unexpected token -> reserved word: then
Syntax error at line 6: unexpected token -> reserved word: then
Syntax error at line 7: unexpected token -> ID, name = fact
```

Por otra parte, algo del comportamiento del analizador sintáctico de TINY es razonable. Por ejemplo, un signo de punto y coma perdido (más que uno sobrante) generará sólo un mensaje de error, y el analizador sintáctico seguirá construyendo el árbol sintáctico correcto como si el signo de punto y coma hubiera estado allí desde el principio, con lo que está realizando una forma rudimentaria de corrección de error en este único caso. Este comportamiento resulta de dos hechos de la codificación. El primero es que el procedimiento `match` no consume un token, lo que da como resultado un comportamiento que es idéntico al de insertar

un token perdido. El segundo es que el procedimiento **stmt_sequence** se escribió de manera que conecte tanto del árbol sintáctico como sea posible en el caso de un error. En particular, se debe tener cuidado para asegurar que los apuntadores hermanos estén conectados dondequiera que se encuentre un apuntador no nulo (los procedimientos del analizador sintáctico están diseñados para devolver un apuntador de árbol sintáctico nulo si se encuentra un error). También, la manera obvia de escribir el cuerpo de **stmt_sequence** basado en el EBNF

```
statement();
while (token==SEMI)
{ match(SEMI);
  statement();
}
```

se puede escribir con una prueba de bucle más complicada:

```
statement();
while ((token!=ENDFILE) && (token!=END) &&
       (token!=ELSE) && (token!=UNTIL))
{ match(SEMI);
  statement();
}
```

El lector puede advertir que los cuatro tokens en esta prueba negativa comprenden el conjunto Siguiente para *stmt-sequence* (*secuencia-sent*). Esto no es un accidente, ya que una prueba puede buscar un token en el conjunto Primero [como lo hacen los procedimientos para *statement* (*sentencia*) y *factor*], o bien, buscar un token que *no* esté en el conjunto Siguiente. Esto último es particularmente efectivo en la recuperación de errores, puesto que si se pierde un símbolo Primero, el análisis sintáctico se detendría. Dejamos para los ejercicios un esbozo del comportamiento del programa para mostrar que un signo de punto y coma perdido en realidad provocaría que el resto del programa se omitiera si **stmt_sequence** se escribiera en la primera forma dada.

Finalmente, advertimos que el analizador sintáctico también se escribió de una manera tal que no puede caer en un bucle infinito cuando encuentra errores (el lector debería haberse preocupado por esto cuando advirtió que **match** no consume un token no esperado). Esto se debe a que, en una ruta arbitraria a través de los procedimientos de análisis sintáctico, con el tiempo se debe encontrar el caso predeterminado de **statement** o **factor**, y ambos *consumen* un token mientras generan un mensaje de error.