



# 4º Ingeniería Informática

## II26 Procesadores de lenguaje

### Análisis semántico

#### Esquema del tema

1. Introducción
2. Esquemas de traducción dirigidos por la sintaxis
3. El árbol de sintaxis abstracta
4. Comprobaciones semánticas
5. Interpretación
6. Introducción a metacomp
7. Algunas aplicaciones
8. Resumen del tema

#### 1. Introducción

Hay determinadas características de los lenguajes de programación que no pueden ser modeladas mediante gramáticas incontextuales y que es necesario comprobar en una fase posterior al análisis sintáctico. Por otro lado, las fases posteriores de la compilación o interpretación necesitan una representación de la entrada que les permita llevar a cabo sus funciones de manera adecuada.

Estas dos vertientes —detección de errores y representación de la información— están muy relacionadas y se solapan en la práctica. Supongamos, por ejemplo, que nuestro lenguaje permite asignaciones según la regla

$$\langle \text{Asignación} \rangle \rightarrow \text{id} := \langle \text{Expresión} \rangle ;$$

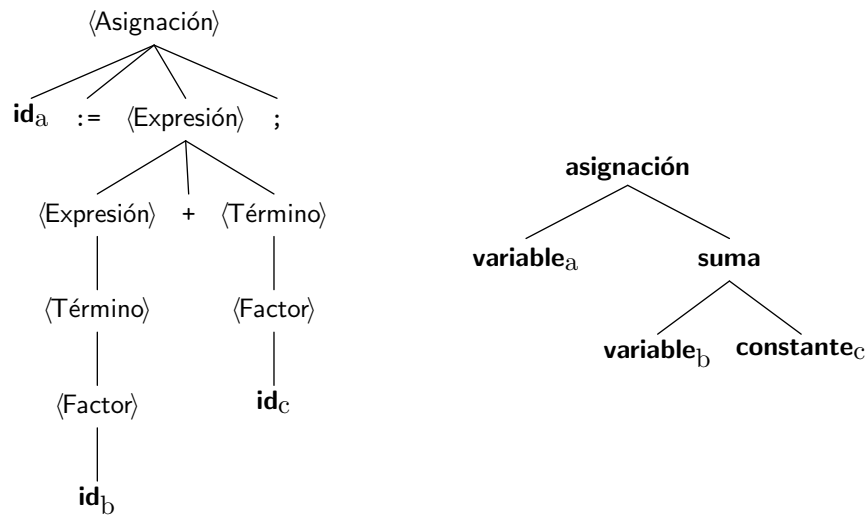
Es habitual que se impongan ciertas restricciones. En nuestro caso, estas podrían ser:

- El identificador de la parte izquierda debe estar declarado previamente.
- El tipo de la expresión debe ser compatible con el del identificador.

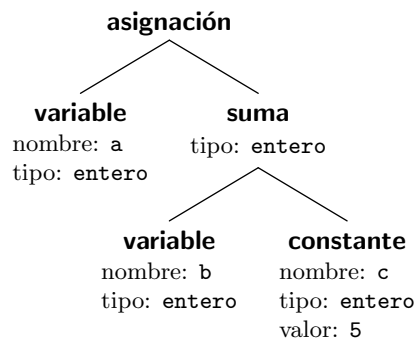
El analizador semántico deberá comprobar que estas dos restricciones se cumplen antes de declarar que la sentencia de asignación está bien formada. Pero sucede que la información necesaria para comprobarlas es útil también para generar código. Esto quiere decir que si tuviéramos una separación estricta entre las fases del compilador, para generar código deberíamos volver a mirar el identificador para saber a qué objeto (variable, función, constante, etc.) corresponde y qué tipo tiene y también deberíamos volver a comprobar los tipos de la expresión para generar el código adecuado.

Por otro lado, aunque en teoría el árbol de análisis sería suficiente para fases posteriores de la compilación o interpretación, es una representación que contiene mucha información redundante. Así, el árbol correspondiente a  $a := b + c$ ; bien podría tener el aspecto del árbol de la izquierda, cuando el de la derecha contiene esencialmente la misma información y resulta más cómodo para

trabajar:



El segundo árbol se conoce como *árbol de sintaxis abstracta* (o AST, de las iniciales en inglés). Como hemos comentado, durante el análisis semántico se recoge una serie de informaciones que resultan de utilidad para fases posteriores. Estas informaciones se pueden almacenar en el árbol, “decorándolo”:



Así el objetivo de la fase de análisis semántico será doble: por un lado detectaremos errores que no se han detectado en fases previas y por otro lado obtendremos el AST decorado de la entrada.

Para ello utilizaremos *esquemas de traducción dirigidos por la sintaxis*, que permitirán asociar acciones a las reglas de la gramática. Estas acciones realizarán comprobaciones y construirán el AST que después se recorrerá para terminar las comprobaciones y será la base para la interpretación o la generación de código.

## 2. Esquemas de traducción dirigidos por la sintaxis

Nuestro objetivo es especificar una serie de acciones que se realizarán durante el análisis de la entrada y tener un mecanismo que nos permita obtener la información necesaria para realizar estas acciones. Para esto añadimos a las gramáticas dos elementos nuevos:

- Acciones intercaladas en las reglas.
- Atributos asociados a los no terminales de la gramática.

### 2.1. Acciones intercaladas en las reglas

Supongamos que tenemos el no terminal `<A>` que deriva dos partes diferenciadas y queremos que se escriba el mensaje “Cambio de parte” tras analizarse la primera. Podemos describir esto

de forma sencilla si aumentamos nuestra gramática incluyendo la acción en la parte derecha de la regla de  $\langle A \rangle$ :

$$\langle A \rangle \rightarrow \langle \text{Parte1} \rangle \{ \text{escribe}(\text{"Cambio de parte"}); \} \langle \text{Parte2} \rangle$$

Podemos entender esta regla extendida como la receta: “analiza la primera parte, una vez encontrada, escribe el mensaje y después analiza la segunda parte”.

## 2.2. Atributos asociados a los no terminales

Si nuestras acciones se limitaran a escribir mensajes que indicaran la fase del análisis en la que nos encontramos, no necesitaríamos mucho más. En general, queremos que las acciones respondan al contenido de los programas que escribimos. Para ello, vamos a añadir a los no terminales una serie de *atributos*. Estos no son más que una generalización de los atributos que tienen los terminales. Vamos a ver un ejemplo.

Supongamos que en un lenguaje de programación se exige que en las subrutinas aparezca un identificador en la cabecera que debe coincidir con el que aparece al final. Tenemos el siguiente fragmento de la gramática del lenguaje:

$$\begin{aligned} \langle \text{Subrutina} \rangle &\rightarrow \langle \text{Cabecera} \rangle \langle \text{Cuerpo} \rangle \langle \text{Fin} \rangle \\ \langle \text{Cabecera} \rangle &\rightarrow \text{subrutina id}(\langle \text{Parámetros} \rangle); \\ \langle \text{Fin} \rangle &\rightarrow \text{fin id} \end{aligned}$$

En la regla correspondiente a  $\langle \text{Fin} \rangle$  queremos comprobar que el identificador es el mismo que en la cabecera. Vamos a definir un atributo del no terminal  $\langle \text{Fin} \rangle$  que será el que nos diga el nombre de la función:

$$\langle \text{Fin} \rangle \rightarrow \text{fin id} \{ \text{si id.lexema} \neq \langle \text{Fin} \rangle.\text{nombre} \text{ entonces error fin si} \}$$

El atributo nombre es lo que se conoce como un *atributo heredado*: su valor se calculará en las reglas en las que  $\langle \text{Fin} \rangle$  aparezca en la parte derecha y se podrá utilizar en las reglas en las que  $\langle \text{Fin} \rangle$  aparezca en la izquierda; será información que “heredará” de su entorno. Fíjate en cómo hemos empleado un atributo de **id** para hacer comprobaciones. El analizador semántico es el que emplea la información contenida en los atributos de los componentes léxicos. Recuerda que el sintáctico sólo ve las etiquetas de las categorías. En cuanto a *error*, suponemos que representa las acciones necesarias para tratar el error.

Ahora tenemos que completar las acciones de modo que  $\langle \text{Fin} \rangle$  tenga algún valor para el nombre. Podemos añadir una acción a la regla de  $\langle \text{Subrutina} \rangle$  para calcular el valor de  $\langle \text{Fin} \rangle$ .nombre:

$$\langle \text{Subrutina} \rangle \rightarrow \langle \text{Cabecera} \rangle \langle \text{Cuerpo} \rangle \{ \langle \text{Fin} \rangle.\text{nombre} := \langle \text{Cabecera} \rangle.\text{nombre} \} \langle \text{Fin} \rangle$$

Lo que hacemos es copiar el atributo nombre que nos “devolverá”  $\langle \text{Cabecera} \rangle$ . No debes confundir el atributo nombre de  $\langle \text{Fin} \rangle$  con el de  $\langle \text{Cabecera} \rangle$ , son completamente independientes. De hecho, el atributo nombre de  $\langle \text{Cabecera} \rangle$  es un *atributo sintetizado*: su valor se calculará en las reglas en las que  $\langle \text{Cabecera} \rangle$  aparezca en la parte izquierda y se utilizará en las reglas donde  $\langle \text{Cabecera} \rangle$  aparezca en la parte derecha; es información que  $\langle \text{Cabecera} \rangle$  “sintetiza” para su entorno. Debemos calcular el valor de nombre en la regla de  $\langle \text{Cabecera} \rangle$ :

$$\langle \text{Cabecera} \rangle \rightarrow \text{subrutina id}(\langle \text{Parámetros} \rangle); \{ \langle \text{Cabecera} \rangle.\text{nombre} := \text{id.lexema} \}$$

## 2.3. Otros elementos de las acciones

Como habrás comprobado, no hemos definido ningún lenguaje formal para escribir las acciones. Asumiremos que se emplean construcciones corrientes de los algoritmos tales como condicionales

o bucles. También haremos referencia en ellos a subrutinas y a variables. Las variables las interpretaremos como variables locales a la regla que estamos analizando o como variables globales si lo indicamos explícitamente. Así, en:

$$\langle \text{Valor} \rangle \rightarrow \text{opsum} \{ \text{si opsum.lexema} = "+" \text{ entonces signo} := 1 \text{ si no signo} := -1 \text{ fin si} \} \\ \langle \text{Valor} \rangle_1 \{ \langle \text{Valor} \rangle.v := \text{signo} * \langle \text{Valor} \rangle_1.v \}$$

no hay ninguna “interferencia” entre las distintas variables signo que se puedan utilizar en un momento dado. Fíjate también en cómo hemos distinguido mediante un subíndice las dos apariciones del no terminal  $\langle \text{Valor} \rangle$ .

Un recurso muy útil son los atributos que contienen listas. Por ejemplo, para una declaración de variables del tipo:

$$\langle \text{DeclVariables} \rangle \rightarrow \langle \text{Listalds} \rangle : \langle \text{Tipo} \rangle \\ \langle \text{Listalds} \rangle \rightarrow \text{id}, \langle \text{Listalds} \rangle | \text{id}$$

podemos hacer lo siguiente:

$$\langle \text{DeclVariables} \rangle \rightarrow \langle \text{Listalds} \rangle : \langle \text{Tipo} \rangle \\ \{ \text{para } v \text{ en } \langle \text{Listalds} \rangle.l \text{ hacer: declara\_var}(v, \langle \text{Tipo} \rangle.t) \text{ fin para} \} \\ \langle \text{Listalds} \rangle \rightarrow \text{id}, \langle \text{Listalds} \rangle_1 \{ \langle \text{Listalds} \rangle.l := \text{concat}([ \text{id.lexema} ], \langle \text{Listalds} \rangle_1.l) \} \\ \langle \text{Listalds} \rangle \rightarrow \text{id} \{ \langle \text{Listalds} \rangle.l := [ \text{id.lexema} ] \}$$

## 2.4. Recopilación

Con lo que hemos visto, podemos decir que un esquema de traducción consta de:

- Una gramática *incontextual* que le sirve de soporte.
- Un conjunto de *atributos* asociados a los símbolos terminales y no terminales.
- Un conjunto de *acciones* asociadas a las partes derechas de las reglas.

Dividimos los atributos en dos grupos:

- Atributos heredados.
- Atributos sintetizados.

Sea  $\langle A \rangle \rightarrow \langle X_1 \rangle \dots \langle X_n \rangle$  una producción de nuestra gramática. Exigiremos que:

- Las acciones de la regla calculen todos los atributos sintetizados de  $\langle A \rangle$ .
- Las acciones situadas a la izquierda de  $\langle X_i \rangle$  calculen todos los atributos heredados de  $\langle X_i \rangle$ .
- Ninguna acción haga referencia a los atributos sintetizados de los no terminales situados a su derecha en la producción.

Las dos últimas reglas nos permitirán integrar las acciones semánticas en los analizadores descendentes recursivos. Cuando se emplean, se dice que el esquema de traducción está basado en una *gramática L-atribuida*. La L indica que el análisis y evaluación de los atributos se pueden hacer de izquierda a derecha.

## EJERCICIO 1

Las siguientes reglas representan parte de las sentencias estructuradas de un lenguaje de programación:

$$\begin{aligned} \langle \text{Programa} \rangle &\rightarrow \langle \text{Sentencias} \rangle \\ \langle \text{Sentencias} \rangle &\rightarrow \langle \text{Sentencia} \rangle \langle \text{Sentencias} \rangle | \langle \text{Sentencia} \rangle \\ \langle \text{Sentencia} \rangle &\rightarrow \text{mientras } \langle \text{Expresión} \rangle \text{ hacer } \langle \text{Sentencias} \rangle \text{ finmientras} \\ \langle \text{Sentencia} \rangle &\rightarrow \text{si } \langle \text{Expresión} \rangle \text{ entonces } \langle \text{Sentencias} \rangle \text{ sino } \langle \text{Sentencias} \rangle \text{ finsi} \\ \langle \text{Sentencia} \rangle &\rightarrow \text{interrumpir} \\ \langle \text{Sentencia} \rangle &\rightarrow \text{otros} \end{aligned}$$

Añade las reglas necesarias para comprobar que la instrucción **interrumpir** aparece únicamente dentro de un bucle.

## EJERCICIO 2

Sea  $G$  la siguiente gramática:

$$\begin{aligned} \langle A \rangle &\rightarrow \langle B \rangle \langle C \rangle | a \\ \langle B \rangle &\rightarrow \langle A \rangle a \langle B \rangle | \langle C \rangle a \\ \langle C \rangle &\rightarrow a \langle C \rangle | aba \langle C \rangle | a \end{aligned}$$

Añade a  $G$  las reglas semánticas necesarias para que el atributo  $ia$  de  $\langle A \rangle$  contenga el número de  $a$ 's al inicio de la cadena generada. Por ejemplo, dadas las cadenas  $aaaaba$ ,  $abaaaa$  y  $aaa$ , los valores de  $ia$  serían 4, 1 y 3, respectivamente.

Puedes utilizar los atributos adicionales que consideres necesarios, pero ninguna variable global. Además, los atributos que añadas deben ser de tipo entero o lógico.

## 2.5. Algunas cuestiones formales

Una pregunta razonable es si puede el símbolo inicial tener atributos heredados y, si esto es así, de dónde proceden. La respuesta varía según los textos. Los que se oponen, defienden que el significado del programa debe depender únicamente de él. Los que están a favor de los atributos heredados, sostienen que permiten formalizar la entrada de información acerca de, por ejemplo, el entorno donde se ejecutará el programa o las opciones de compilación.

Una situación similar se da con los símbolos terminales: hay autores que piensan que no deberían tener atributos en absoluto; otros defienden que sólo deben tener atributos sintetizados; y los hay que opinan que pueden tener tanto atributos heredados como sintetizados.

Otro aspecto sobre el que hay diferencias de interpretación es el de los efectos laterales de las reglas. En algunos textos, las reglas de evaluación de los atributos no pueden tener efectos laterales. Otros autores defienden que esta distinción no es más que un problema de implementación ya que es posible añadir un nuevo atributo que represente el entorno e ir actualizándolo adecuadamente. Esto únicamente supone una incomodidad, pero no un problema formal. Esta es la posición que seguiremos nosotros, permitiendo que haya efectos laterales en el cálculo de atributos.

Quizá los ejemplos más claros de existencia de efectos laterales sean el manejo de la tabla de símbolos y el control de errores. Cuando se encuentra una declaración en un programa es necesario actualizar la tabla de símbolos de modo que sea posible reconocer ocurrencias posteriores del identificador. Podemos suponer que existe una tabla de símbolos global o tener un atributo que lleve copias de la tabla de un sitio a otro.

Por otro lado, en caso de que se encuentre un error, es necesario tomar medidas como detener la generación de código. Una manera sencilla de marcar esta circunstancia es tener una variable

global de tipo lógico que indique si se ha encontrado algún error. Nuevamente, sería posible, pero no necesario, tener un atributo que transporte esa información.

## 2.6. Eliminación de recursividad por la izquierda y atributos

En el tema de análisis sintáctico vimos cómo se pueden modificar las producciones con recursividad por la izquierda para lograr que la gramática sea LL(1). El problema con las transformaciones es que dejan una gramática que tiene muy poca relación con la original, lo que hace que escribir los atributos y las reglas correspondientes sea difícil sobre la gramática transformada. Sin embargo, se pueden escribir los atributos sobre la gramática original y después convertirlos de una manera bastante mecánica.

Como las GPDRs no suelen necesitar recursividad por la izquierda, es fácil que no tengas que utilizar esta transformación. Pese a todo, vamos a ver cómo se hace la transformación sobre un ejemplo. Partimos del siguiente esquema de traducción:

$$\begin{aligned} \langle E \rangle &\rightarrow \langle E \rangle_1 + \langle T \rangle \{ \langle E \rangle.v := \langle E \rangle_1.v + \langle T \rangle.v \} \\ \langle E \rangle &\rightarrow \langle T \rangle \{ \langle E \rangle.v := \langle T \rangle.v \} \\ \langle T \rangle &\rightarrow \text{num} \{ \langle T \rangle.v := \text{num}.v \} \end{aligned}$$

Comenzamos por transformar la gramática:

$$\begin{aligned} \langle E \rangle &\rightarrow \langle T \rangle \langle E' \rangle \\ \langle E' \rangle &\rightarrow + \langle T \rangle \langle E' \rangle \\ \langle E' \rangle &\rightarrow \lambda \\ \langle T \rangle &\rightarrow \text{num} \end{aligned}$$

Como vemos, el problema es que cuando vemos el sumando derecho en  $\langle E' \rangle$ , no tenemos acceso al izquierdo. La idea será crear un atributo heredado que nos diga el valor del sumando izquierdo. En la primera regla lo podemos calcular directamente. En la segunda regla, realizamos la suma y la transmitimos:

$$\begin{aligned} \langle E \rangle &\rightarrow \langle T \rangle \{ \langle E' \rangle.h := \langle T \rangle.v \} \langle E' \rangle \\ \langle E' \rangle &\rightarrow + \langle T \rangle \{ \langle E' \rangle_1.h := \langle E' \rangle.h + \langle T \rangle.v \} \langle E' \rangle_1 \end{aligned}$$

Con esto, el atributo  $h$  contendrá siempre la suma, que es el valor que tenemos que devolver finalmente. Por eso hay que “transmitir” el nuevo valor al atributo  $v$ :

$$\begin{aligned} \langle E \rangle &\rightarrow \langle T \rangle \{ \langle E' \rangle.h := \langle T \rangle.v \} \langle E' \rangle \{ \langle E \rangle.v := \langle E' \rangle.v \} \\ \langle E' \rangle &\rightarrow + \langle T \rangle \{ \langle E' \rangle_1.h := \langle E' \rangle.h + \langle T \rangle.v \} \langle E' \rangle_1 \{ \langle E' \rangle.v := \langle E' \rangle_1.v \} \end{aligned}$$

¿Qué hacemos cuando  $\langle E' \rangle$  se reescribe como la cadena vacía? En este caso, basta con devolver como sintetizado el valor que se hereda. El esquema de traducción completo queda:

$$\begin{aligned} \langle E \rangle &\rightarrow \langle T \rangle \{ \langle E' \rangle.h := \langle T \rangle.v \} \langle E' \rangle \{ \langle E \rangle.v := \langle E' \rangle.v \} \\ \langle E' \rangle &\rightarrow + \langle T \rangle \{ \langle E' \rangle_1.h := \langle E' \rangle.h + \langle T \rangle.v \} \langle E' \rangle_1 \{ \langle E' \rangle.v := \langle E' \rangle_1.v \} \\ \langle E' \rangle &\rightarrow \lambda \{ \langle E' \rangle.v := \langle E' \rangle.h \} \\ \langle T \rangle &\rightarrow \text{num} \{ \langle T \rangle.v := \text{num}.v \} \end{aligned}$$

### EJERCICIO 3

Escribe el árbol de análisis de  $3+4$  con el esquema original y el transformado. Decóralos.

## EJERCICIO 4

Transforma el siguiente esquema de traducción para eliminar la recursividad por la izquierda:

$$\begin{aligned}
 \langle E \rangle &\rightarrow \langle E \rangle_1 + \langle T \rangle \{ \langle E \rangle.v := \langle E \rangle_1.v + \langle T \rangle.v \} \\
 \langle E \rangle &\rightarrow \langle T \rangle \{ \langle E \rangle.v := \langle T \rangle.v \} \\
 \langle T \rangle &\rightarrow \langle T \rangle_1 * \langle F \rangle \{ \langle T \rangle.v := \langle T \rangle_1.v * \langle F \rangle.v \} \\
 \langle T \rangle &\rightarrow \langle F \rangle \{ \langle T \rangle.v := \langle F \rangle.v \} \\
 \langle F \rangle &\rightarrow \langle E \rangle \{ \langle F \rangle.v := \langle E \rangle.v \} \\
 \langle F \rangle &\rightarrow \text{num} \{ \langle F \rangle.v := \text{num}.v \}
 \end{aligned}$$

El siguiente ejercicio presenta la transformación de una manera más general.

## EJERCICIO\* 5

Si tenemos las siguientes reglas en un esquema de traducción

$$\begin{aligned}
 \langle A \rangle &\rightarrow \langle A \rangle_1 \langle Y \rangle \{ \langle A \rangle.a := g(\langle A \rangle_1.a, \langle Y \rangle.y) \} \\
 \langle A \rangle &\rightarrow \langle X \rangle \{ \langle A \rangle.a := f(\langle X \rangle.x) \}
 \end{aligned}$$

podemos transformarlas en

$$\begin{aligned}
 \langle A \rangle &\rightarrow \langle X \rangle \{ \langle A \rangle.h := f(\langle X \rangle.x) \} \langle A \rangle \{ \langle A \rangle.a := \langle A \rangle.s \} \\
 \langle A \rangle &\rightarrow \langle Y \rangle \{ \langle A \rangle_1.h := g(\langle A \rangle.h, \langle Y \rangle.y) \} \langle A \rangle_1 \{ \langle A \rangle.s := \langle A \rangle_1.s \} \\
 \langle A \rangle &\rightarrow \lambda \{ \langle A \rangle.s := \langle A \rangle.h \}
 \end{aligned}$$

Comprueba que la transformación es correcta analizando  $\langle X \rangle \langle Y \rangle_1 \langle Y \rangle_2$  mediante las dos versiones y comparando el valor de a.

## 2.7. Implementación de los esquemas de traducción

La interpretación de las acciones como sentencias que se ejecutan al pasar el análisis por ellas permite implementar los esquemas de traducción de manera sencilla. Para ello se modifica la implementación del analizador recursivo descendente correspondiente a la gramática original de la siguiente manera:

- Los atributos heredados del no terminal  $\langle A \rangle$  se interpretan como parámetros de entrada de la función `Analiza_A`.
- Los atributos sintetizados del no terminal  $\langle A \rangle$  se interpretan como parámetros de salida de la función `Analiza_A`.
- Las acciones semánticas, una vez traducidas al lenguaje de programación correspondiente, se insertan en la posición correspondiente según su orden en la parte derecha donde aparecen.

En la práctica, es frecuente que, si el lenguaje de programación (como C) no permite devolver más de un valor, los atributos sintetizados del no terminal se pasen por referencia.

En Python existe una solución bastante cómoda. Comenzamos por definir una clase vacía:

```
class Atributos:
    pass
```

Antes de llamar a una función o método de análisis, creamos un objeto de esta clase y le añadimos los atributos heredados del no terminal. La correspondiente función de análisis creará los atributos sintetizados. Este es el único parámetro que se pasa. Así, la traducción de la regla:

$$\langle E \rangle \rightarrow \langle T \rangle \{ \langle R \rangle . h := \langle T \rangle . v \} \langle R \rangle \{ \langle E \rangle . v := \langle R \rangle . v \}$$

es la siguiente:

```
def analiza_E(E):
    T= Atributos() # Atributos de T
    R= Atributos() # Atributos de R
    analiza_T(T)
    R.h= T.v # Creamos un atributo heredado de R
    analiza_R(R)
    E.v= R.v # Creamos un atributo sintetizado de E
```

Lógicamente, tendríamos que haber añadido el código de control de errores.

## 2.8. Atributos en GPDR

La interpretación que hemos hecho de los esquemas de traducción se traslada de forma natural a las GPDR. Por ejemplo, la traducción de:

$$\langle E \rangle \rightarrow \langle T \rangle_1 \{ \langle E \rangle . v := \langle T \rangle_1 . v \} (+ \langle T \rangle_2 \{ \langle E \rangle . v := \langle E \rangle . v + \langle T \rangle_2 . v \})^*$$

es simplemente:

```
def analiza_E(E):
    T1= Atributos() # Atributos de T1
    T2= Atributos() # Atributos de T2
    analiza_T(T1)
    E.v= T1.v
    while token.cat=="suma":
        token= alex.siguiete()
        analiza_T(T2)
        E.v= E.v+T2.v
```

Como antes, habría que añadir el correspondiente código para controlar errores.

## 3. El árbol de sintaxis abstracta

Como hemos comentado en la introducción, una de las posibles representaciones semánticas de la entrada es el árbol de sintaxis abstracta o AST.

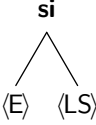
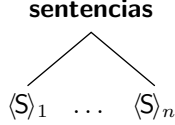


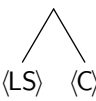
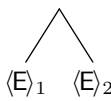
Aunque es similar a los árboles de análisis sintáctico, tiene algunas diferencias importantes:

- No aparecen todos los componentes léxicos del programa. Por ejemplo:
  - No es necesario incluir los paréntesis de las expresiones.
  - No se necesitan los separadores o terminadores de las sentencias.
  - ...
- Pueden aparecer otros componentes no estrictamente sintácticos, como acciones de coerción de tipos.



### 3.1. Construcción

Para construir los árboles, debemos comenzar por definir qué elementos emplearemos para cada estructura:

Estructura	Representación	Estructura	Representación
<b>si</b> if $\langle E \rangle$ then $\langle LS \rangle$ end		<b>sentencias</b> begin $\langle S \rangle_1 \dots \langle S \rangle_n$ end	
<b>mientras</b> while $\langle C \rangle$ do $\langle LS \rangle$ end		<b>asignación</b> id := $\langle E \rangle$ ;	
<b>repetir</b> repeat $\langle LS \rangle$ until $\langle C \rangle$ ;		<b>suma</b> $\langle E \rangle_1 + \langle E \rangle_2$	
...	...	...	...

Fíjate que el árbol resultante puede representar programas en diversos lenguajes de programación del estilo C, Pascal, etc.

Ahora debemos utilizar los atributos para construir el árbol. Utilizando el atributo arb para devolver el árbol que construye cada no terminal, podemos hacer algo parecido a:

```

(Sentencia) → if <Expresión> then <Sentencias> end
              {<Sentencia>.arb:= NodoSi(<Expresión>.arb, <Sentencias>.arb)}
(Sentencia) → while <Expresión> do <Sentencias> end
              {<Sentencia>.arb:= NodoMientras(<Expresión>.arb, <Sentencias>.arb)}
(Sentencia) → repeat <Sentencias> until <Expresión> ;
              {<Sentencia>.arb:= NodoRepetir(<Sentencias>.arb, <Expresión>.arb)}
(Sentencia) → id:= <Expresión> ;
              {<Sentencia>.arb:= NodoAsignación(id.lexema, <Expresión>.arb)}

```

Para la implementación hay dos opciones principales:

- Utilizar funciones (NodoSi, NodoMientras, ...) que devuelvan una estructura de datos adecuada.
- Utilizar objetos (NodoSi, NodoMientras, ...) para cada uno de los nodos.

Probablemente, la segunda opción sea la más cómoda para el trabajo posterior con el árbol.

En cuanto a la lista de sentencias, podemos utilizar nodos que tengan un grado variable, para eso almacenamos una lista con los hijos:

$$\langle \text{Sentencias} \rangle \rightarrow \{l:=\lambda\} (\langle \text{Sentencia} \rangle \{l:=l+\langle \text{Sentencia} \rangle.\text{arb}\})^* \{ \langle \text{Sentencias} \rangle.\text{arb}:= \text{NodoSentencias}(l) \}$$

El tratamiento de las expresiones es sencillo. Por ejemplo, para la suma podemos hacer:

$$\langle \text{Expresión} \rangle \rightarrow \langle \text{Término} \rangle_1 \{ \text{arb}:=\langle \text{Término} \rangle_1.\text{arb} \} ( + \langle \text{Término} \rangle_2 \{ \text{arb}:=\text{NodoSuma}(\text{arb}, \langle \text{Término} \rangle_2.\text{arb}) \})^* \{ \langle \text{Expresión} \rangle.\text{arb}:= \text{arb} \}$$

Las restas, productos, etc, se tratarían de forma similar. Puedes comprobar que de esta manera el AST resultante respeta la asociatividad por la izquierda.

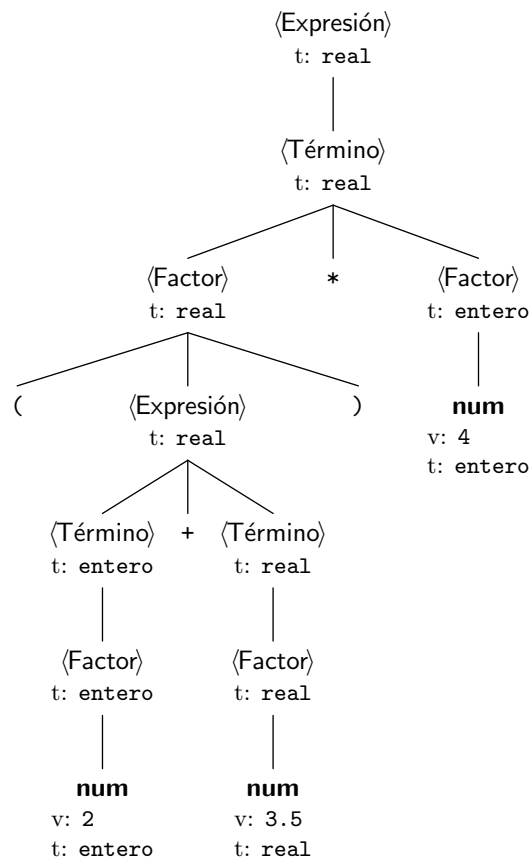
#### EJERCICIO\* 6

En el caso de la asociatividad por la derecha tenemos dos opciones: crear una lista de operandos y recorrerla en orden inverso para construir el árbol o utilizar recursividad por la derecha, que no da problemas para el análisis LL(1). Utiliza ambas posibilidades para escribir sendos esquemas de traducción que construyan los AST para expresiones formadas por sumas y potencias de identificadores siguiendo las reglas habituales de prioridad y asociatividad.

### 3.2. Evaluación de atributos sobre el AST

Un aspecto interesante del AST es que se puede utilizar para evaluar los atributos sobre él, en lugar de sobre la gramática inicial. Si reflexionas sobre ello, es lógico. Todo el proceso de evaluación de los atributos se puede ver como el etiquetado de un árbol (el árbol de análisis), pero no hay nada que impida que el árbol sobre el que se realiza la evaluación sea el AST.

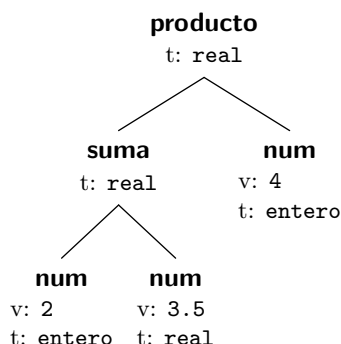
Un ejemplo sería el cálculo de tipos. Si tenemos la expresión  $(2+3.5)*4$ , podemos calcular los tipos sobre el árbol de análisis:



Para realizar esta evaluación, podríamos utilizar una gramática similar a la siguiente<sup>1</sup>:

$$\begin{aligned} \langle \text{Expresión} \rangle &\rightarrow \langle \text{Término} \rangle_1 \{t:= \langle \text{Término} \rangle_1.t\} (+ \langle \text{Término} \rangle_2 \{t:= \text{másgeneral}(t, \langle \text{Término} \rangle_2.t)\})^* \\ &\quad \{ \langle \text{Expresión} \rangle.t:= t \} \\ \dots & \\ \langle \text{Término} \rangle &\rightarrow \langle \text{Factor} \rangle_1 \{t:= \langle \text{Factor} \rangle_1.t\} (* \langle \text{Factor} \rangle_2 \{t:= \text{másgeneral}(t, \langle \text{Factor} \rangle_2.t)\})^* \\ &\quad \{ \langle \text{Término} \rangle.t:= t \} \\ \dots & \\ \langle \text{Factor} \rangle &\rightarrow \text{num} \{ \langle \text{Factor} \rangle.t:= \text{num}.t \} \\ \langle \text{Factor} \rangle &\rightarrow (\langle \text{Expresión} \rangle) \{ \langle \text{Factor} \rangle.t:= \langle \text{Expresión} \rangle.t \} \\ \dots & \end{aligned}$$

También podemos realizar el cálculo sobre el AST:



Esto se haría junto con el resto de comprobaciones semánticas.

La elección acerca de si evaluar atributos en el AST o durante el análisis descendente es básicamente una cuestión de simplicidad. Generalmente, podemos decir que los atributos sintetizados tienen una dificultad similar en ambos casos. Sin embargo, cuando la gramática ha sufrido transformaciones o hay muchos atributos heredados, suele ser más fácil evaluarlos sobre el AST.

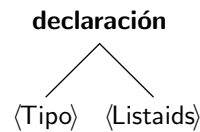
Otro aspecto interesante a la hora de decidir qué atributos evaluar sobre el árbol de análisis y cuáles sobre el AST está en los propios nodos del árbol. Supongamos que queremos tener nodos diferentes para la suma entera y la suma real. En este caso, tendremos que comprobar los tipos directamente sobre el árbol de análisis (o crear un AST y después modificarlo, pero esto puede ser forzarlo demasiado).

## 4. Comprobaciones semánticas

Los atributos nos permitirán llevar a cabo las comprobaciones semánticas que necesite el lenguaje. En algunos casos, utilizaremos los atributos directamente (posiblemente evaluados sobre el AST), por ejemplo en la comprobación que hacíamos de que el identificador al final de la función era el mismo que al principio.

En otros casos, los atributos se utilizan indirectamente mediante estructuras globales, por ejemplo la tabla de símbolos. Un ejemplo sería la comprobación de que un identificador no se ha declarado dos veces. Si hemos utilizado un nodo similar a:

<sup>1</sup>Utilizamos la función `másgeneral` que devuelve el tipo más general de los que se le pasan como parámetros (el tipo real se considera más general que el entero).



el método de comprobación sería:

**Objeto** NodoDeclaración:

```

...
Método compsemánticas()
...
para  $i \in \text{listaid}$  hacer
    si TablaSímbolos.existe( $i$ ) entonces
        error
    si no
        TablaSímbolos.inserta( $i$ , tipo)
    fin si
fin para
...
fin compsemánticas
...
fin NodoDeclaración
  
```

Utilizando directamente el esquema de traducción habría que duplicar parte del código:

```

<Declaración>  → <Tipo> { <Listaid>.t := <Tipo>.t } <Listaid>
<Listaid>      → id, { <Listaid>1.t := <Listaid>.t } <Listaid>1
                { si TablaSímbolos.existe(id.lexema)
                  entonces error
                  si no TablaSímbolos.inserta(id.lexema, <Listaid>.t) }
<Listaid>      → id
                { si TablaSímbolos.existe(id.lexema)
                  entonces error
                  si no TablaSímbolos.inserta(id.lexema, <Listaid>.t) }
  
```

#### 4.1. La tabla de símbolos

Durante la construcción del AST, las comprobaciones semánticas y, probablemente, durante la interpretación y la generación de código necesitaremos obtener información asociada a los distintos identificadores presentes en el programa. La estructura de datos que permite almacenar y recuperar esta información es la tabla de símbolos.

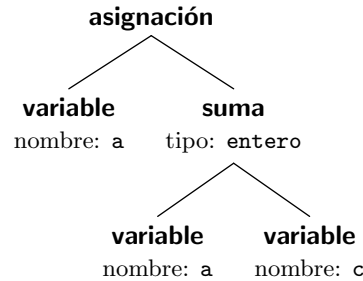
En principio, la tabla debe ofrecer operaciones para:

- Insertar información relativa a un identificador.
- Recuperar la información a partir del identificador.

La estructura que puede realizar estas operaciones de manera eficiente es la tabla *hash* (que en Python está disponible mediante los diccionarios). La implementación habitual es una tabla que asocia cada identificador a información tal como su naturaleza (constante, variable, nombre de función, etc.), su tipo (entero, real, booleano, etc.), su valor (en el caso de constantes), su dirección (en el caso de variables y funciones), etc.

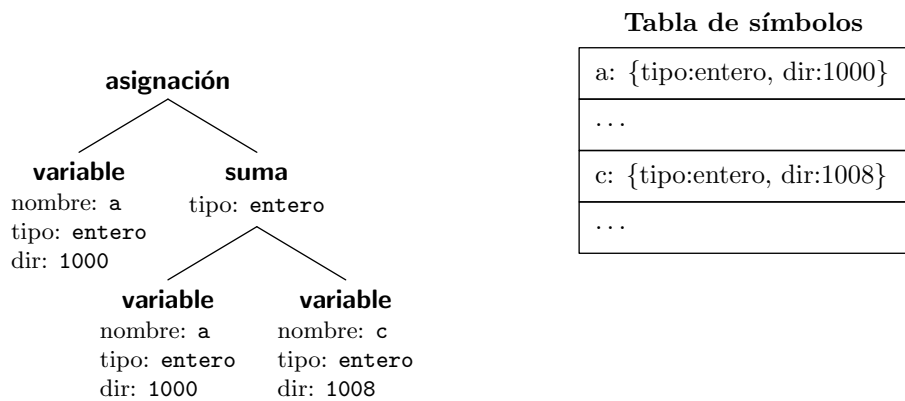
Es importante tener únicamente una tabla; si tenemos varias, por ejemplo, una para constantes y otra para variables, el acceso se hace más difícil ya que para comprobar las propiedades de un identificador hay que hacer varias consultas en lugar de una.

Una cuestión importante es cómo se relacionan la tabla y el AST. Tenemos distintas posibilidades, que explicaremos sobre el siguiente árbol, correspondiente a la sentencia `a= a+c`:

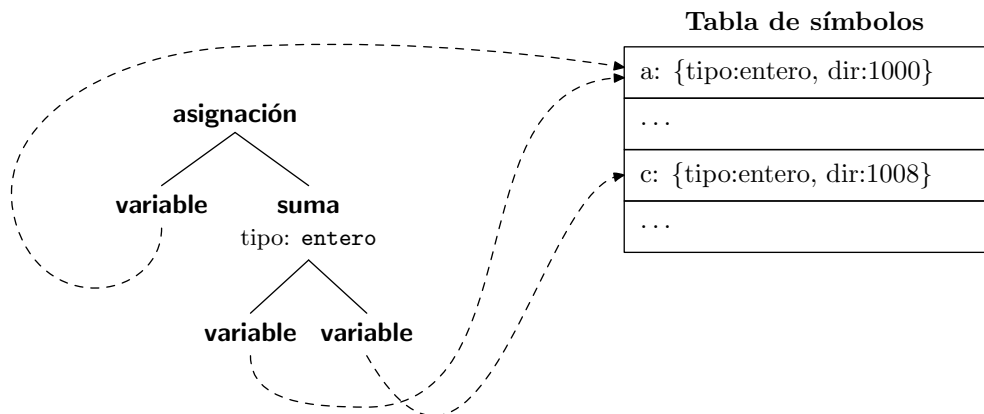


La primera posibilidad es dejar el árbol tal cual y cada vez que necesitemos alguna información, por ejemplo el valor de `a`, consultar la tabla. Esta opción sería la más adecuada para situaciones en las que se vaya a recorrer el árbol pocas veces, por ejemplo en una calculadora donde se evalúe cada expresión una sola vez.

Otra posibilidad es ir decorando cada una de las hojas con toda la información que se vaya recopilando. Por ejemplo, si durante el análisis averiguamos el tipo y dirección de las variables, pasaríamos a almacenar la nueva información:

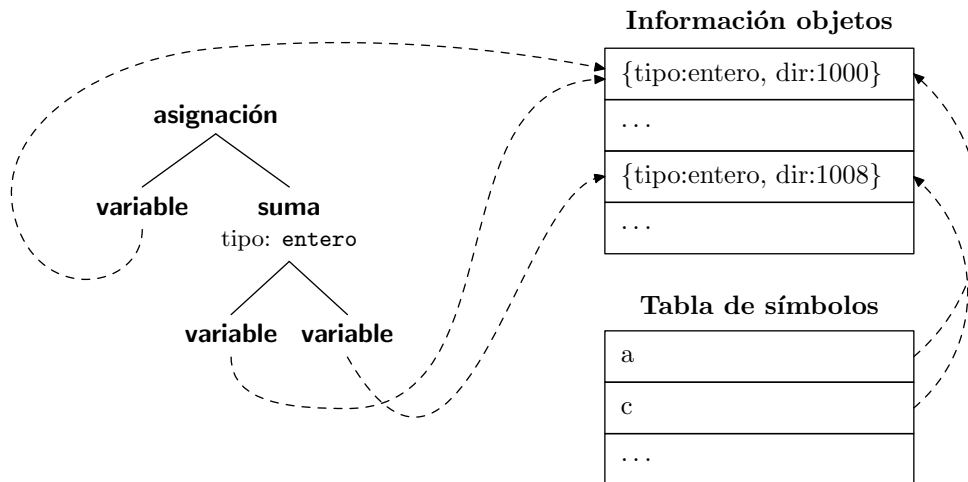


Esto tiene costes muy elevados en tiempo —tendremos que actualizar repetidamente un número de nodos que puede ser elevado— y espacio —hay mucha información que se almacena repetida—. Podemos reducir estos costes guardando en cada nodo un puntero a la entrada correspondiente en la tabla:



Finalmente, la opción más adecuada cuando tenemos lenguajes que presenten ámbitos anidados (veremos en otro tema cómo representar los ámbitos en la tabla) es guardar la información sobre

los objetos en otra estructura de datos a la que apunta tanto la tabla como los nodos del árbol:



### 4.2. Comprobaciones de tipos

Un aspecto importante del análisis semántico es la comprobación de los tipos de las expresiones. Esta comprobación se hace con un doble objetivo:

- Detectar posibles errores.
- Averiguar el operador o función correcto en casos de sobrecarga y polimorfismo.

Para representar los tipos en el compilador, se emplean lo que se denominan *expresiones de tipo* (ET). La definición de estas expresiones es generalmente recursiva. Un ejemplo sería:

- Los tipos básicos: real, entero, . . . , además de los tipos especiales *error\_de\_tipo* y ausencia de tipo (*void*) son ETs.
- Si  $n_1$  y  $n_2$  son enteros,  $rango(n_1, n_2)$  es una ET.
- Si  $T$  es una ET, también lo es  $puntero(T)$ .
- Si  $T_1$  y  $T_2$  son ETs, también lo es  $T_1 \rightarrow T_2$ .
- Si  $T_1$  y  $T_2$  son ETs, también lo es  $vector(T_1, T_2)$ .
- Si  $T_1$  y  $T_2$  son ETs, también lo es  $T_1 \times T_2$ .
- Si  $T_1, \dots, T_k$  son ETs y  $N_1, \dots, N_k$  son nombres de campos,  $registro((N_1, T_1), \dots, (N_k, T_k))$  es una ET.

Ten en cuenta que según los lenguajes, estas definiciones cambian. Es más, los lenguajes pueden restringir qué expresiones realmente resultan en tipos válidos para los programas (por ejemplo, en Pascal no se puede definir un vector de funciones).

Algunos ejemplos de declaraciones y sus expresiones correspondientes serían:

Declaración	Expresión de tipo
<code>int v[10];</code>	$vector(rango(0, 9), entero)$
<code>struct {</code> <code>  int a;</code> <code>  float b;</code> <code>} st;</code>	$registro((a, entero), (b, real))$
<code>int *p;</code>	$puntero(entero)$
<code>int f(int, char);</code>	$entero \times car\acute{a}cter \rightarrow entero$
<code>void f2(float);</code>	$real \rightarrow void$

### 4.2.1. Equivalencia de tipos

Comprobar si dos expresiones de tipo,  $T_1$  y  $T_2$ , son equivalentes es muy sencillo. Si ambas corresponden a tipos elementales, son equivalentes si son iguales. En caso de que correspondan a tipos estructurados, hay que comprobar si son el mismo tipo y si los componentes son equivalentes.

En muchos lenguajes se permite dar nombre a los tipos. Esto introduce una sutileza a la hora de comprobar la equivalencia. La cuestión es, dada una declaración como

```
typedef int a;
typedef int b;
```

¿son equivalentes a y b? La respuesta depende del lenguaje (o, en casos como el Pascal, de la implementación). Existen dos criterios para la equivalencia:

**Equivalencia de nombre:** dos expresiones de tipo con nombre son equivalentes si y sólo si tienen el mismo nombre.

**Equivalencia estructural:** dos expresiones de tipo son equivalentes si y sólo si tienen la misma estructura.

Hay argumentos a favor de uno y otro criterio. En cualquier caso, hay que tener en cuenta que para comprobar la equivalencia estructural ya no sirve el método trivial presentado antes. En este caso, puede haber ciclos, lo que obliga a utilizar algoritmos más complicados.

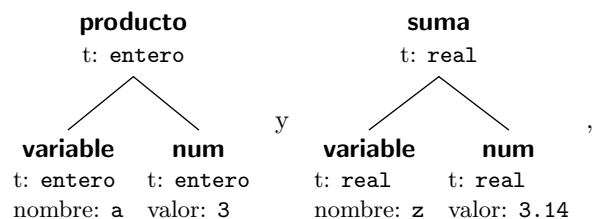
### 4.2.2. Comprobación de tipos en expresiones

Para comprobar los tipos en las expresiones podemos utilizar un atributo que indique el tipo de la expresión. Este tipo se infiere a partir de los distintos componentes de la expresión y de las reglas de tipos del lenguaje.

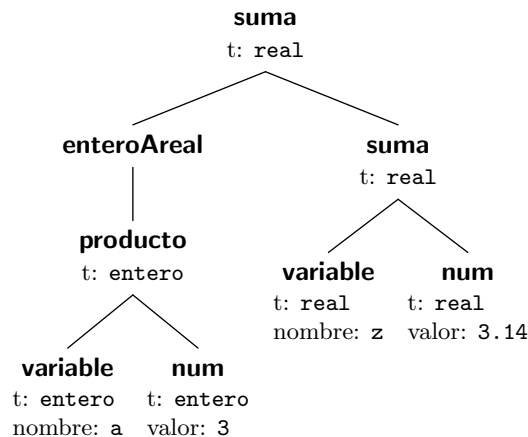
A la hora de calcular el tipo hay varias opciones: podemos calcularlo sobre la gramática, sobre el AST o al construir el AST. En cualquier caso, puede ser útil tener una función similar a la de la sección 3.2 que refleje las peculiaridades de los tipos del lenguaje. Por ejemplo, si tenemos tipos entero y real con las reglas habituales de promoción, esta función devolvería los resultados según esta tabla:

<i>Primero</i>	<i>Segundo</i>	<i>Resultado</i>
<i>entero</i>	<i>entero</i>	<i>entero</i>
<i>entero</i>	<i>real</i>	<i>real</i>
<i>real</i>	<i>entero</i>	<i>real</i>
<i>real</i>	<i>real</i>	<i>real</i>
	<i>otros</i>	<i>error_tipo</i>

El cálculo de atributos de tipo debería ser trivial sobre la gramática o sobre el AST. En cuanto al cálculo al construir el AST, se puede, por ejemplo, hacer que el constructor compruebe los tipos de los componentes que recibe. En este caso, si es necesario, se puede añadir un nodo de promoción de tipos. Por ejemplo, si vamos a crear un nodo suma a partir de los árboles:



podemos añadir un nodo intermedio para indicar la promoción:



Respecto a la propagación de los errores de tipo, hay que intentar evitar un exceso de mensajes de error. Para ello se pueden utilizar distintas estrategias:

- Se puede hacer que *error\_tipo* sea compatible con cualquier otro tipo.
- Se puede intentar inferir cuál sería el tipo en caso de no haber error. Por ejemplo: si el operador en que se ha detectado el error es el y-lógico, se puede devolver como tipo el tipo lógico.

## 5. Interpretación

La utilización del AST hace que la interpretación sea bastante sencilla. Una vez lo hemos construido, tenemos dos aproximaciones para la interpretación:

- Podemos representar la entrada mediante una serie de instrucciones similares a un lenguaje máquina con algunas características de alto nivel. Este es el caso, por ejemplo, de Java y el Java bytecode. Esta aproximación representa prácticamente una compilación.
- La otra alternativa es representar la entrada de manera similar al AST y recorrerlo para ejecutar el programa.

Utilizaremos la segunda opción. Los nodos correspondientes a las expresiones tendrán un método al que llamaremos *evalúa* y que devolverá el resultado de evaluar el subárbol correspondiente. Por ejemplo, en el nodo de la suma tendríamos:

**Objeto** NodoSuma:

```

...
Método evalúa()
    devuelve i.evalúa() + d.evalúa(); // i y d son los hijos del nodo
fin evalúa
...
fin NodoSuma
  
```

Aquellos nodos que representen sentencias tendrán el método *interpreta*. Por ejemplo, para un nodo que represente un mientras, el método correspondiente sería:

**Objeto** NodoMientras:

```

...
Método interpreta()
    mientras condición.evalúa()=cierto:
        sentencias.interpreta()
    fin mientras
fin interpreta
...
fin NodoMientras
  
```



Las variables se representarían mediante entradas en una tabla (que, en casos sencillos, puede ser la tabla de símbolos) que contenga el valor correspondiente en cada momento. Una asignación consiste simplemente en evaluar la expresión correspondiente y cambiar la entrada de la tabla.

La ejecución del programa consistirá en una llamada al método *interpreta* del nodo raíz del programa. Alternativamente, si hemos guardado el programa como una lista de sentencias, la ejecución consistirá en un bucle que recorra la lista haciendo las llamadas correspondientes.

Lógicamente, existen otras maneras de recorrer el árbol. En particular, si no lo hemos representado con objetos, podemos recorrerlo mediante funciones recursivas o de manera iterativa. En este último caso, podemos emplear una pila auxiliar o ir “enhebrando” el árbol durante su construcción. Esta es probablemente la opción más rápida con una implementación cuidadosa, pero es también la que más esfuerzo exige (especialmente si hay que recorrer el árbol en distintos órdenes).

## 6. Introducción a metacomp

Mediante **metacomp** se pueden traducir esquemas de traducción a programas Python que implementan los correspondientes analizadores descendentes recursivos.

Un programa **metacomp** se divide en varias secciones. Cada sección se separa de la siguiente por un carácter “%” que ocupa la primera posición de una línea. La primera sección contiene la especificación del analizador léxico. Las secciones pares contienen código de usuario. La tercera y sucesivas secciones impares contienen el esquema de traducción. Es decir, un programa **metacomp** se escribe en un fichero de texto siguiendo este formato:

```
Especificación léxica
%
Código de usuario
%
Esquema de traducción
%
Código de usuario
%
Esquema de traducción
:
```

Las secciones de “código de usuario” se utilizan para declarar estructuras de datos y variables, definir funciones e importar módulos útiles para el procesador de lenguaje que estamos especificando. Estos elementos deberán codificarse en Python. La herramienta **metacomp** copia literalmente estos fragmentos de código en el programa que produce como salida.

### 6.1. Especificación léxica

La especificación léxica consiste en una serie de líneas con tres partes cada una:

- Un nombre de categoría o la palabra **None** si la categoría se omite.
- Un nombre de función de tratamiento o **None** si no es necesario ningún tratamiento.
- Una expresión regular.

Los analizadores léxicos generados por **metacomp** dividen la entrada siguiendo la estrategia avariciosa y, en caso de conflicto, prefiriendo las categorías que aparecen en primer lugar. Por cada lexema encontrado en la entrada se llama a la correspondiente función de tratamiento, que normalmente se limita a calcular algún atributo adicional a los tres que tienen por defecto todos los componentes: *lexema*, *nlinea* y *cat*, que representan, respectivamente, el lexema, el número de línea y la etiqueta de categoría del componente.

Un ejemplo de especificación léxica para una calculadora sencilla sería:

categoria	expresión regular	acciones	atributos
num	[0-9] <sup>+</sup>	calcular valor emitir	valor
sum	[ <sup>-</sup> +]	copiar lexema emitir	lexema
abre	\(	emitir	
cierra	\)	emitir	
espacio	[\t] <sup>+</sup>	omitir	
nl	\n	emitir	

En `metacomp`, lo podemos escribir así:

```
num    valor [0-9]+
sum    None  [-+ ]
abre   None  \(
cierra None  \)
None   None  [ \t ]+
nl     None  \n
```

Donde `valor` sería una función como la siguiente:

```
def valor(componente):
    componente.v= int(componente.lexema)
```

No necesitamos ningún tratamiento para `sum` ya que `metacomp` añade por defecto el atributo `lexema` (que es el que hemos usado para calcular el valor de los enteros).

## 6.2. Esquema de traducción

Las secciones de esquema de traducción contienen las reglas de la GPDR utilizando una sintaxis muy similar a la de la asignatura. Cada regla tiene una parte izquierda, un símbolo  $\rightarrow$ , una parte derecha y un punto y coma. Los no terminales se escriben marcados mediante  $\langle$  y  $\rangle$ . Los terminales tienen que coincidir con los declarados en la especificación léxica<sup>2</sup>. Se pueden emplear los operadores regulares de concatenación (implícita), disyunción (mediante  $|$ ), opcionalidad (mediante  $?$ ), clausura (mediante  $*$ ) y clausura positiva (mediante  $+$ ). Por ejemplo, la regla

$$\langle E \rangle \rightarrow \langle T \rangle (\text{sum} \langle T \rangle)^*$$

se escribiría en `metacomp` así:

```
<E> -> <T> ( sum <T> )*;
```

Las acciones de las reglas se escriben encerradas entre arrobas (`@`) y sin fin de línea entre ellas. Para poder referirse a los terminales y no terminales, estos están numerados por su orden de aparición en la parte derecha de la regla. El no terminal de la izquierda no tiene número y, si no hay ambigüedad, la primera aparición de cada símbolo no necesita número. Si tenemos las acciones de la regla anterior:

$$\langle E \rangle \rightarrow \langle T \rangle_1 \{v:= \langle T \rangle_1.v\} \\ (\text{sum} \langle T \rangle_2 \{\text{si sum.lexema} = \text{"+" entonces } v:= v+\langle T \rangle_2.v \text{ si no } v:= v-\langle T \rangle_2.v \text{ fin si}\})^* \\ \{\langle E \rangle.v:= v\}$$

podemos escribirlas en `metacomp` así:

<sup>2</sup>Existe también la posibilidad de utilizar una sintaxis especial para categorías con un sólo lexema, pero no lo comentaremos.

```

<E> -> <T> @v= T.v@
      ( sum <T>
        @if sum.lexema=="+":@
        @ v+= T2.v@
        @else:@
        @ v-= T2.v@
      )*
      @E.v= v@;

```

## 7. Algunas aplicaciones

Vamos a aprovechar `metacomp` para reescribir los ejemplos que vimos en el tema de análisis léxico: la suma de las columnas de un fichero con campos separados por tabuladores y la lectura de ficheros de configuración.

### 7.1. Ficheros organizados por columnas

La representación en `metacomp` de la especificación léxica podría ser:

```

1  numero    valor    [0-9]+
2  separacion None    \t
3  nl        None    \n
4  None      None    [ ]+

```

La función `valor` se define en la zona de auxiliares junto con dos funciones para tratamiento de errores:

```

5  %
6  def valor(componente):
7      componente.v= int(componente.lexema)
8
9  def error(linea, mensaje):
10     sys.stderr.write("Error en línea %d: %s\n" % (linea, mensaje))
11     sys.exit(1)
12
13 def error_lexico(linea, cadena):
14     if len(cadena)> 1:
15         error(linea, "no he podido analizar la cadena %s." % repr(cadena))
16     else:
17         error(linea, "no he podido analizar el carácter %s." % repr(cadena))
18
19 ncol= 0

```

La función `error` la utilizamos para escribir mensajes genéricos. La función `error_lexico` es llamada por `metacomp` cuando encuentra un error léxico. Los dos parámetros que recibe son la línea donde se ha producido el error y el carácter o caracteres que no se han podido analizar.

Podemos emplear el siguiente esquema de traducción:

```

<Fichero> → {suma:=0}(<Línea> nl {suma:= suma+<Línea>.v}) * {<Fichero>.suma:= suma}
<Línea> → numero1 {global ncol; col:=1; si col= ncol entonces <Línea>.v:= numero.v fin si}
          (separación número2
           {col:= col+1; si col= ncol entonces <Línea>.v:= numero2.v fin si}
          )*

```

Hemos supuesto que el símbolo inicial tiene un atributo sintetizado que leerá el entorno. Además, la columna deseada está en la variable global *ncol*. Representamos esto en *metacomp* así:

```

20  %
21  <Fichero>->  @suma= 0@
22                ( <Linea> @suma+= Linea.v@ nl )*
23                @Fichero.suma= suma@ ;
24
25  <Fichero>-> error
26                @error(mc_al.linea(), "línea mal formada")@
27                ;
28
29  <Linea>-> numero
30                @col= 1@
31                @if col== ncol:@
32                @ Linea.v= numero.v@
33                ( separacion numero
34                @col+= 1@
35                @if col== ncol:@
36                @ Linea.v= numero2.v@
37                )*
38                @if col< ncol:@
39                @ error(numero.nlinea, "no hay suficientes columnas")@
40                ;
41

```

La regla `<Fichero>-> error` nos permite capturar los errores que se produzcan durante el análisis. Hay varias maneras de tratar los errores, pero hemos optado por abortar el programa. Hemos utilizado *mc\_al*, que contiene el analizador léxico, para averiguar el número de línea del error.

Finalmente, hemos decidido que el programa tenga un parámetro que indique qué columna sumar. Si no se especifica ningún parámetro adicional, se leerá la entrada estándar. En caso de que haya varios parámetros, se sumará la columna deseada de cada uno de los ficheros especificados, escribiéndose el total. Con esto, el programa principal es:

```

42  %
43  def main():
44      global ncol
45      if len(sys.argv)== 1:
46          sys.stderr.write("Error, necesito un número de columna.\n")
47          sys.exit(1)
48      try:
49          ncol= int(sys.argv[1])
50      except:
51          sys.stderr.write("Error, número de columna mal formado.\n")
52          sys.exit(1)
53
54      if len(sys.argv)== 2:
55          A= AnalizadorSintactico(sys.stdin)
56          suma= A.Fichero.suma
57      else:
58          suma= 0
59      for arg in sys.argv[2:]:
60          try:
61              f= open(arg)
62          except:

```

```

63         sys.stderr.write("Error, no he podido abrir el fichero %s.\n" % arg)
64         sys.exit(1)
65         A= AnalizadorSintactico(f)
66         suma+= A.Fichero.suma
67         print "La suma es %d." % suma

```

Por defecto, `metacomp` genera una función `main` que analiza `sys.stdin`. Si queremos modificar ese comportamiento, debemos crear nuestra propia función. Como ves, la mayor parte del código se dedica a analizar posibles errores en los parámetros. La construcción más interesante está en las dos líneas que crean el analizador. La asignación

```
A= AnalizadorSintactico(f)
```

hace que se cree un analizador que inmediatamente analizará el fichero `f` (que debe estar abierto). Cuando termina el análisis, el objeto `A` tiene un atributo con el nombre del símbolo inicial y que tiene sus atributos sintetizados. Así es como transmitimos el resultado.

## 7.2. Ficheros de configuración

Vamos ahora a programar el módulo de lectura de los ficheros de configuración. Haremos que el resultado del análisis sea un diccionario en el que las claves sean las variables y los valores asociados sean los leídos del fichero.

El analizador léxico se escribiría así:

```

1  variable  None    [a-zA-Z][a-zA-Z0-9]*
2  igual     None    =
3  valor     None    [0-9]+|\("[^"\n]*\"
4  nl        None    (#[^\n]*)?\n
5  None      None    [ \t]+

```

Los únicos auxiliares que necesitamos son las funciones de tratamiento de errores:

```

6  %
7  def error(linea, mensaje):
8      sys.stderr.write("Error en línea %d: %s\n" % (linea, mensaje))
9      sys.exit(1)
10
11 def error_lexico(linea, cadena):
12     if len(cadena)> 1:
13         error(linea, "no he podido analizar la cadena %s." % repr(cadena))
14     else:
15         error(linea, "no he podido analizar el carácter %s." % repr(cadena))

```

Nuestro esquema de traducción será:

```

<Fichero> → {dic:= {}}(((<Línea> {dic[<Línea>.izda]:= <Línea>.dcha}|λ) nl)* {<Fichero>.dic:= dic}
<Línea> → variable {<Línea>.izda:= variable.lexema } igual valor {<Línea>.dcha:= valor.lexema}

```

Pasado a `metacomp`:

```

16  %
17  <Fichero>-> @dic= {}@
18              ( (<Linea> @dic[Linea.izda]= Linea.dcha@)? nl )*
19              @Fichero.dic= dic@ ;
20
21  <Fichero>-> error

```

```

22             @error(mc_al.linea(), "línea mal formada")@
23             ;
24
25 <Linea>-> variable
26             @Linea.izda= variable.lexema@
27             igual valor
28             @Linea.dcha= valor.lexema@
29             ;
30

```

Como ves, hemos utilizado el operador de opcionalidad para representar la disyunción  $((\text{Linea})|\lambda)$ .

## 8. Resumen del tema

- El analizador semántico tiene dos objetivos:
  - Hacer comprobaciones que no se hagan durante el análisis léxico o sintáctico.
  - Crear una representación adecuada para fases posteriores.
- Implementaremos el análisis semántico en dos partes:
  - Mediante esquemas de traducción dirigidos por la sintaxis.
  - Recorriendo el AST.
- Un esquema de traducción dirigido por la sintaxis añade a las gramáticas:
  - Acciones intercaladas en las partes derechas de las reglas.
  - Atributos asociados a los no terminales.
- Dos tipos de atributos: heredados y sintetizados.
- Las acciones deben garantizar que se evalúan correctamente los atributos.
- Se pueden implementar los esquemas de traducción sobre los analizadores sintácticos interpretando los atributos como parámetros y añadiendo el código de las acciones al código del analizador.
- El cálculo de algunos atributos y algunas comprobaciones semánticas son más fáciles sobre el AST.
- La tabla de símbolos se puede implementar eficientemente mediante una tabla *hash*.
- Las comprobaciones de tipos se complican si se introducen nombres. Dos criterios:
  - Equivalencia de nombre.
  - Equivalencia estructural.
- La interpretación se puede realizar mediante recorridos del AST.
- `metacomp` permite implementar cómodamente esquemas de traducción dirigidos por la sintaxis.