Compiladores
Facultad de Ciencias UNAM
Análisis Semáctico: *Gramáticas con Atributos*

Lourdes Del Carmen González Huesca [*]

9 de abril de 2024

Resumen

En este punto se ha estudiado una línea de acciones a tomar para definir si un programa dado es suceptible de traducirle a lenguaje de bajo nivel o no, para finalmente ejecutarse. Un análisis importante es el análisis sintáctico que determina si un programa o cadena es generado a partir de una gramática dada. Este análisis se hace a través de encontrar una derivación que genere a la cadena y de manera complementaria un árbol de sintaxis concreta o parse-tree. Otro análisis importante que no se ha considerado hasta ahora es el semántico, el cual es un tipo de análisis que decreta si un programa tiene sentido o no, por ejemplo al verificar que los tipos de argumentos de una función coincidan con los esperados en su definición. Este análisis semántico se hace a través de complementar las gramáticas con reglas llamadas atributos. Este análisis se sirve del árbol generado en el análisis sintáctico y de la tabla de símbolos ya que puede agregar etiquetas en los nodos que aporten información sobre la semántica del programa o agregar información a la tabla de símbolos.

Términos clave:
Es: Gramáticas con atributos, Atributo Sintético, Atributo heredado, Gramática L-Atribuida, Gramática S-Atribuida.
En: Attributed grammar, Synthesized Attribute, Inherited Attribute, L-Attributed Grammar, S-Attributed grammar.

El analizador semático corresponde a la última fase del compilador en la parte de front-end. Este analizador verificará la corrección del programa más alla de la forma, es decir a nivel semántico, respecto a la definición del lenguaje de programación en particular.

Recordemos que un lenguaje de programación se define mediante:

1. Reglas sintácticas, para describir los símbolos válidos y usando gramáticas libres de contexto.
2. Reglas semánticas, que definen características estáticas y dinámicas relacionadas con el significado de los símbolos.

Estas definiciones siguen un estilo declarativo al sólo establecer reglas que definen las especificaciones del lenguaje, inclusive se puede usar una descripción con lenguaje natural para definir las estructuras y el funcionamiento de ellas en el lenguaje de programación.

Una gramática libre de contexto no especifica cómo debe ser *parseado* un programa sino que define las reglas de construcción y esta parte sintáctica es el fundamento para los análisis léxico y sintáctico.

[*]Material revisado por el servicio social de Apoyo a la Docencia y Asesoría Académica en la Facultad de Ciencias de la UNAM con clave 2023-12/12-292, de Diana Laura Nicolás Pavia.

De las reglas semánticas, las estáticas establecen propiedades de los programas que serán verificadas en tiempo de compilación para obtener código que preservará estas propiedades. Y las reglas dinámicas permitirán describir cómo se ejecuta un programa. Ambas reglas semánticas son de gran utilidad en esta fase de compilación ya que cada compilador tiene diferentes reglas semánticas particulares que deben asegurar que las estructuras y valores definidos en el lenguaje se cumplan. Las reglas semánticas del compilador corresponden a un tipo especial de gramáticas, que también siguen un estilo declarativo ya que no especifican el orden en que deben ser aplicadas ni mucho menos indican el tipo de información que es sensible para estudiar la corrección del programa más allá de su forma.

Después del análisis sintáctico se debe considerar información importante que permita indagar si el programa tiene sentido, esto es al analizar en alto nivel su "significado". El significado real de un programa se estudia a través la **semántica denotacional** que permite asociar un objeto matemático a cada construcción sintáctica del lenguaje a través de funciones semánticas que describen el efecto de ejecutar dicha construcción. Este análisis no se estudiará en este curso pero se puede consultar [3].

La información importante o relevante son detalles de las partes de un programa que no pueden obtenerse del análisis sintáctico, por ejemplo verificar que los tipos de argumentos de una función coincidan con los esperados por su definición. Esta información ha sido parcialmente recuperada del programa por el lexer y el parser y, almacenada en la tabla de símbolos pero es necesario que sea incluida en el árbol sintáctico [1].

Decimos que el análisis semántico puede ser descrito en téminos de anotaciones o **decoraciones** en el árbol sintáctico. Las anotaciones son llamadas **atributos**. Los atributos y su relación con los tokens fueron obtenidos por el analizador léxico. Recordemos que un token es una palabra significativa que consta de dos partes, el nombre y su atributo: la primera es la representación del tipo de unidad léxica y la segunda es el valor de dicha unidad.

En el diseño de un compilador es importante definir cuáles atributos **no** son libres de contexto ya que éstos serán los valores que aparecen en el programa, que no están generados por una gramática libre de contexto y que por lo tanto no son resultado de la fase anterior del compilador es decir del analizador sintáctico o parser. Estos atributos deben ser incluidos en el árbol sintáctico para poder complementar una representación intermedia que sea fiel al programa original y susceptible de ser traducido a un lenguaje de bajo nivel.
Los resultados de la fase del análisis dependiente del contexto constan de

- anotar el árbol sintáctico con atributos, por ejemplo al usar apuntadores a identificadores en la tabla de símbolos;
- recorrer el árbol para generar una representación intermedia alternativa que describa el control de flujo del programa.

En esta parte estudiaremos cómo obtener la información sensible al contexto que aclare definiciones (por ejemplo, los tipos de argumentos, variables globales, etc.) además de algunos requerimientos para la ejecución. Por otra parte, si se desea estudiar más a fondo las propiedades de programas que sean dependientes de la ejecución se pueden abordar los métodos formales. Éstos son herramientas y formalismos que aseguran que se cumplen propiedades y requerimientos en tiempo de ejecución al relacionar el diseño y la implementación de especificaciones, por ejemplo las pruebas unitarias, la Lógica de Hoare, el análisis sintáctico, los sistemas de tipos, lenguajes con tipos dependientes, etc. De ellos, sólo nos ocuparemos de los sistemas de tipos para describir propiedades de seguridad de programas en tiempo de compilación. Estos sistemas los veremos en una nota posterior.

[1] En esta fase nos referimos al árbol sintáctico al resultado del análisis sintáctico ya sea como un árbol de sintaxis concreta o una simplificación del mismo mediante un árbol de sintaxis abstracta.

Gramáticas con atributos

Una gramática con atributos es un complemento a una gramática libre de contexto al anotar las producciones con atributos. Estas gramáticas no especifican el significado del programa sólo ayudan a asociarlo con valores que explican su significado, es así que para cada símbolo de la gramática A, se le asocia una función que describe su valor $A.\texttt{val}$ o algún otro atributo:

- para los símbolos no-terminales, el valor es generado por la parte derecha de las producciones al dar significado o valor a la cadena de tokens derivada del símbolo
- los símbolos terminales que tengan atributo se le asocia el valor que depende del programa

El valor particular que se genere en cada regla de producción de una gramática libre de contexto clasifica a la regla en:

Reglas de copiado
El atributo es copia de algún otro atributo, por ejemplo en las producciones unitarias de la forma $A \to B$:

$$A \to B \quad \triangleright \quad A.\texttt{val} := B.\texttt{val}$$

Reglas con función semántica
El atributo es calculado con funciones específicas dirigidas por el diseño del lenguaje y cuyos argumentos son atributos de la parte derecha de la producción:

$$A \to \alpha \quad \triangleright \quad A.\texttt{val} := \mathcal{F}(\alpha)$$

Obsérvese que no se puede hacer referencia a valores o atributos fuera de una producción y que cada símbolo de la gramática puede tener cero o más atributos. La forma en que se calculan los atributos permite clasificarlos en:

Atributo sintético obtiene su valor de un enunciado hacia la izquierda en una producción. Los símbolos terminales tienen propiedades intrínsecas y es por esto que son atributos sintéticos, obtienen su valor del programa original a través de la información recabada en la tabla de símbolos desde el lexer.

Atributo heredado obtiene su valor cuando el mismo símbolo no-terminal está a la derecha de la producción o al usar valores o atributos de otros símbolos. Es decir que la información contextual en el árbol sintáctico debe fluir de un símbolo en un nodo superior o en el mismo nivel. De esta forma, las reglas de producción de la gramática de atributos heredados pueden ser usadas muchas veces para obtener diferentes valores dependiendo del contexto y que se va heredando desde la información que está almacenada en la tabla de símbolos.

Una gramática con atributos está bien definida si las reglas determinan un conjunto único de valores para cada árbol sintáctico derivado de la gramática del lenguaje. Y es no-circular si nunca genera ciclos en el árbol sintáctico al calcular el flujo de los atributos.

Existen tres **esquemas de traducción** para aplicar las reglas de la gramática con atributos y obtener el valor de los atributos:

1. Esquema Inadvertido (*oblivious*): no toma en cuenta un orden de las reglas de producción e ignora el árbol sintáctico generado por el análisis anterior. Se escoge un orden conveniente para calcular los atributos y se repite este orden para calcular todos los atributos.
2. Esquema Dinámico: este método toma en cuenta la forma del parse tree al calcular una gráfica de dependencias. El orden para calcular los atributos de los nodos está determinado por el orden topológico en el árbol.

3. Esquema Estático: se realiza un análisis de las reglas semánticas en la construcción del compilador y se establece un orden en cada regla de producción para obtener los atributos en ella. Así se calculan todos los atributos dependiendo de la regla.

Una gráfica de dependencias en este caso establece el flujo de la información entre las instancias de los atributos en el parse tree. Las aristas son las restricciones de las reglas semánticas y los nodos son los diferentes atributos asociados a cada símbolo. Esta gráfica es particular e independiente a cada parse tree.

Tipos de gramáticas con atributos

Decimos que una gramática es **S-atribuida** si todos sus atributos son sintéticos y los argumentos de las funciones semánticas usan únicamente símbolos en la parte derecha de la producción. El resultado de un atributo es la parte izquierda de la producción. Estas gramáticas están asociadas a los parsers LR, es decir que el cálculo de los atributos se realiza desde las hojas y hacia la raíz que es la misma forma en que se obtiene el parse tree. Este tipo de gramática favorece que los atributos puedan ser calculados al vuelo en la fase de análisis sintáctico.

Por otro lado, decimos que una gramática es **L-atribuida** si los atributos de los nodos son evaluados al visitar los nodos del parse tree de una sola vez de izquierda a derecha (**L**eft-to-right) y en un recorrido a profundidad:

- cada atributo sintético a la izquierda depende de los heredados o de la parte derecha
- cada atributo heredado a la derecha depende de los atributos heredados de la izquierda o de los atributos de la parte izquierda

Estas gramáticas están asociadas a los parsers LL ya que los atributos pueden ser calculados desde la raíz y hacia las hojas.

Toda gramática S-atribuida es L-atribuida pero no al revés. El tipo gramática S-atribuida es la más general y es la que en la práctica se implementa junto con un parser LR. Si el parser es LL entonces se implementará una L-atribuida.

Los atributos decoran el árbol sintáctico y se almacenan en los nodos; esto puede realizarse de dos formas:

1. Después del análisis sintáctico y como resultado del análisis semántico; es claro que las fases del compilador están separadas y se implementa una gramática con atributos en especial junto con un esquema de traducción apropiado.

2. Al mismo tiempo que se realiza el análisis sintáctico; es decir que la gramática tenga intercaladas funciones que permitan decorar el parse tree y generar una representación intermedia al mismo tiempo, entonces el tipo de compilador es uno *one-pass*.

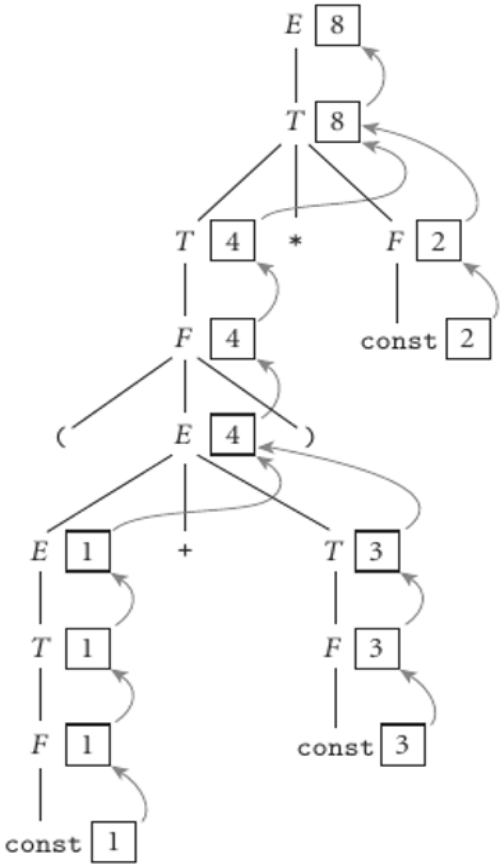
Evaluación de atributos

Calcular los atributos se realiza ya sea desde las hojas o desde la raíz dependiendo del tipo de gramática S-atribuida o L-atribuida. A continuación veremos dos ejemplos, en la presentación correspondiente se encuentran los ejemplos de las mismas gramáticas pero cuyas funciones semánticas construyen árboles de sintaxis abstracta.

Ejemplo 1 (Gramática S-atribuida)**.** La siguente gramática de expresiones aritméticas calcula los atributos desde las hojas

1.	$E_1 \longrightarrow E_2$ + T	✂ E_1.val := sum(E_2.val, T.val)
2.	$E_1 \longrightarrow E_2$ - T	✂ E_1.val := difference(E_2.val, T.val)
3.	$E \longrightarrow T$	✂ E.val := T.val
4.	$T_1 \longrightarrow T_2$ * F	✂ T_1.val := product(T_2.val, F.val)
5.	$T_1 \longrightarrow T_2$ / F	✂ T_1.val := quotient(T_2.val, F.val)
6.	$T \longrightarrow F$	✂ T.val := F.val
7.	$F_1 \longrightarrow$ - F_2	✂ F_1.val := additive_inverse(F_2.val)
8.	$F \longrightarrow$ (E)	✂ F.val := E.val
9.	$F \longrightarrow$ const	✂ F.val := const.val

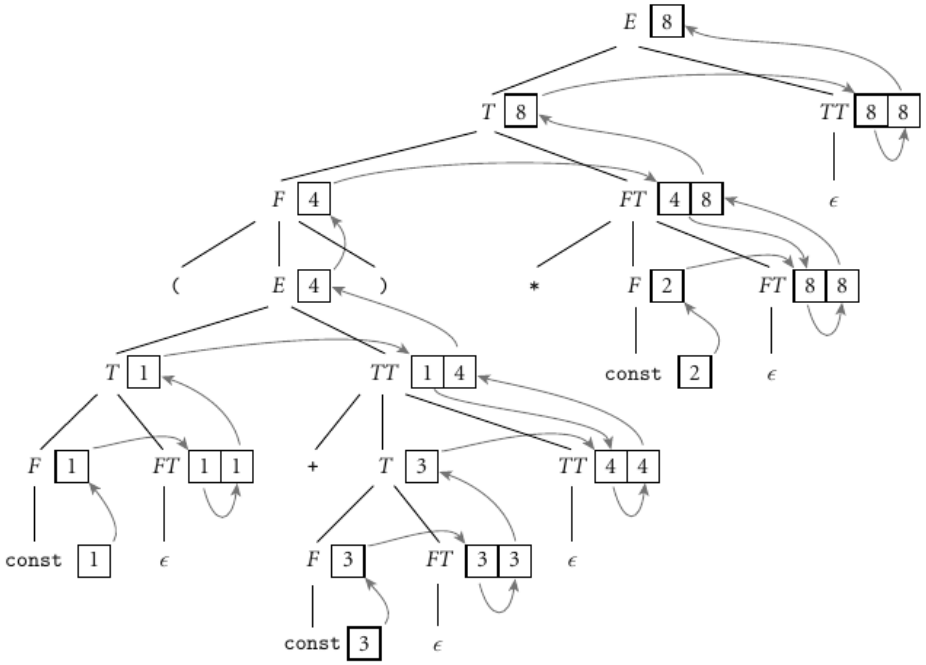
El siguiente árbol corresponde a la expresión $(1 + 3) * 2$, las flechas curvas indican la forma en que se obtuvieron los atributos.



Ejemplo 2 (Gramática L-atribuida)**.** La siguente gramática de expresiones aritméticas calcula los atributos desde la raíz siguiendo una estrategia a profundidad y de izquierda a derecha

1. $E \longrightarrow T\ TT$
   ✂ TT.st := T.val ✂ E.val := TT.val
2. $TT_1 \longrightarrow +\ T\ TT_2$
   ✂ TT_2.st := TT_1.st + T.val ✂ TT_1.val := TT_2.val
3. $TT_1 \longrightarrow -\ T\ TT_2$
   ✂ TT_2.st := TT_1.st − T.val ✂ TT_1.val := TT_2.val
4. $TT \longrightarrow \epsilon$
   ✂ TT.val := TT.st
5. $T \longrightarrow F\ FT$
   ✂ FT.st := F.val ✂ T.val := FT.val
6. $FT_1 \longrightarrow *\ F\ FT_2$
   ✂ FT_2.st := FT_1.st × F.val ✂ FT_1.val := FT_2.val
7. $FT_1 \longrightarrow /\ F\ FT_2$
   ✂ FT_2.st := FT_1.st ÷ F.val ✂ FT_1.val := FT_2.val
8. $FT \longrightarrow \epsilon$
   ✂ FT.val := FT.st
9. $F_1 \longrightarrow -\ F_2$
   ✂ F_1.val := − F_2.val
10. $F \longrightarrow (\ E\ )$
    ✂ F.val := E.val
11. $F \longrightarrow$ `const`
    ✂ F.val := const.val

El siguiente árbol corresponde a la expresión $(1 + 3) * 2$, las flechas curvas indican la forma en que se obtuvieron los atributos.



Ejercicios

1. Considera la siguiente gramática cuyas funciones permiten obtener tamaños de tipos básicos y arreglos:

T	$\rightarrow$	BC	$t = B.\texttt{type} \quad w = B.\texttt{width}$ $T.\texttt{type} = C.\texttt{type} \quad T.\texttt{width} = C.\texttt{width}$
B	$\rightarrow$	int	$B.\texttt{type} = integer$ $B.\texttt{width} = 4$
B	$\rightarrow$	float	$B.\texttt{type} = float$ $B.\texttt{width} = 8$
C	$\rightarrow$	ε	$C.\texttt{type} = t$ $C.\texttt{width} = w$
C	$\rightarrow$	$[\text{num}]C_1$	$C.\texttt{type} = array(\text{num}.value, C_1.type)$ $C.\texttt{width} = \text{num}.value \times C_1.\texttt{width}$

   Los atributos `type` y `width` de los símbolos no-terminales son sintetizados mientras que las variables t y w se usan para pasar la información del tipo y el tamaño (en bytes) de un nodo B al nodo de la producción $C \rightarrow \varepsilon$ en el árbol de sintaxis concreta. Estos últimos atributos son heredados para la variable C.

   a) Describe con palabras qué realizan las funciones semánticas para el resto de las producciones.

   b) Muestra el árbol de sintaxis decorado para la expresión `float[4][2][3]` indicando el flujo del cálculo de ellos usando flechas en el árbol.

2. Considera la siguiente gramática libre de contexto para constantes de tipo flotante (no incluye la notación exponencial).

   $C \rightarrow digits.digits$

   $digits \rightarrow digit\ more\_digits$

   $more\_digits \rightarrow digits \mid \varepsilon$

   $digit \rightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$

   a) Aumenta las producciones de esta gramática para convertirla en una gramática con atributos en donde las funciones semánticas acumulen el valor de la constante en un atributo `val` en la raíz del parse tree de una expresión. Tu respuesta debe ser una gramática S-atribuida

   b) Calcula el parse tree para la espresión 3.2

3. Considera la siguiente gramática para expresiones aritméticas y que construye un árbol sintáctico usando los atributos `st` y `ptr` que son apuntadores a nodos de un árbol sintáctico.

E	$\rightarrow$	$T\ TT$	$\texttt{TT.st} := \texttt{T.ptr}$
			$\texttt{E.ptr} := \texttt{TT.ptr}$
TT_1	$\rightarrow$	$+\ T\ TT_2$	$\texttt{TT}_2\texttt{.st} := \texttt{make\_bin\_op}(\text{“}+\text{”}, \texttt{TT}_1\texttt{.st}, \texttt{T.ptr})$
			$\texttt{TT}_1\texttt{.ptr} := \texttt{TT}_2\texttt{.ptr}$
TT_1	$\rightarrow$	$-\ T\ TT_2$	$\texttt{TT}_2\texttt{.st} := \texttt{make\_bin\_op}(\text{“}-\text{”}, \texttt{TT}_1\texttt{.st}, \texttt{T.ptr})$
			$\texttt{TT}_1\texttt{.ptr} := \texttt{TT}_2\texttt{.ptr}$
TT	$\rightarrow$	ε	$\texttt{TT.ptr} := \texttt{TT.st}$
T	$\rightarrow$	$F\ FT$	$\texttt{FT.st} := \texttt{F.ptr}$
			$\texttt{T.ptr} := \texttt{FT.ptr}$
FT_1	$\rightarrow$	$*\ F\ FT_2$	$\texttt{FT}_2\texttt{.st} := \texttt{make\_bin\_op}(\text{“}*\text{”}, \texttt{FT}_1\texttt{.st}, \texttt{F.ptr})$
			$\texttt{FT}_1\texttt{.ptr} := \texttt{FT}_2\texttt{.ptr}$
FT_1	$\rightarrow$	$/\ F\ FT_2$	$\texttt{FT}_2\texttt{.st} := \texttt{make\_bin\_op}(\text{“/”}, \texttt{FT}_1\texttt{.st}, \texttt{F.ptr})$
			$\texttt{FT}_1\texttt{.ptr} := \texttt{FT}_2\texttt{.ptr}$
FT	$\rightarrow$	ε	$\texttt{FT.ptr} := \texttt{FT.st}$
F_1	$\rightarrow$	$-\ F_2$	$\texttt{FT.ptr} := \texttt{make\_un\_op}(\text{“}-\text{”}, \texttt{FT}_2\texttt{.ptr})$
F	$\rightarrow$	(E)	$\texttt{F.ptr} := \texttt{E.ptr}$
F	$\rightarrow$	`const`	$\texttt{F.ptr} := \texttt{make\_leaf}(\texttt{const.val})$

a) Muestra los pasos para obtener un árbol de sintaxis para la expresión $(4 * 5)$ indicando el o los subárboles de parsing, los atributos de cada nodo usando flechas para indicar los apuntadores que construyen el árbol.

b) Aumentar la gramática para inicializar los atributos sintéticos en cada nodo del árbol sintáctico para indicar la posición (línea y columna) en la cual aparece el constructor de dicho nodo en el programa fuente. Asumir que en análisis léxico se ha obtenido dicha posición para cada token.

Referencias

[1] Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman. *Compilers, Principles, Techniques and Tools*. Pearson Education Inc., Second edition, 2007.

[2] Jean-Christophe Filliâtre. Curso Compilation (inf564) école Polytechnique, Palaiseau, Francia. `http://www.enseignement.polytechnique.fr/informatique/INF564/`, 2018. Material en francés.

[3] Hanne Riis Nielson and Flemming Nielson. *Semantics with Applications: An Appetizer (Undergraduate Topics in Computer Science)*. Springer-Verlag, Berlin, Heidelberg, 2007.

[4] Frank Pfenning. Notas del curso (15-411) Compiler Design. `https://www.cs.cmu.edu/~fp/courses/15411-f14/`, 2014.

[5] Michael Lee Scott. *Programming Language Pragmatics*. Morgan-Kauffman Publishers, Third edition, 2009.

[6] Yunlin Su and Song Y. Yan. *Principles of Compilers, A New Approach to Compilers Including the Algebraic Method*. Springer-Verlag, Berlin Heidelberg, 2011.

[7] Steve Zdancewic. Notas del curso (CIS 341) - Compilers, Universidad de Pennsylvania, Estados Unidos. `https://www.cis.upenn.edu/~cis341/current/`, 2018.