

Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
IC-5701 Compiladores e intérpretes
Proyecto #2, la fase de análisis contextual de un compilador

Historial de revisiones:

- 2021.10.13,17: Versión base (v0).

Lea con cuidado este documento. Si encuentra errores en el planteamiento¹, por favor comuníquese los inmediatamente al profesor.

Objetivo

Al concluir este proyecto Ud. habrá terminado de comprender los detalles relativos a fase de análisis contextual de un compilador escrito "a mano" al aplicar las técnicas expuestas por Watt y Brown en su libro *Programming Language Processors in Java*. Ud. deberá extender el compilador del lenguaje Δ (escrito en Java), desarrollado por Watt y Brown, de manera que sea capaz de procesar el lenguaje descrito en el enunciado del Proyecto #1 y en la sección *Lenguaje fuente* que aparece abajo. El lenguaje extendido será denominado Δ_{xt} en el resto de este documento. Además de realizar el trabajo de análisis léxico y sintáctico logrado en el Proyecto #1, su compilador debe ser capaz de procesar el lenguaje Δ_{xt} conforme las características contextuales (identificación y tipos) especificadas en este documento. Su compilador deberá coexistir con un ambiente de edición, compilación y ejecución ('IDE'). Se le suministra un IDE construido en Java por el Dr. Luis Leopoldo Pérez, ajustado por estudiantes de Ingeniería en Computación del TEC.

Bases

Para entender las técnicas expuestas en el libro de texto, Ud. deberá estudiar el compilador del lenguaje imperativo Δ , ambos desarrollados en Java por los profesores David Watt y Deryck Brown. El compilador ha sido ubicado en la carpeta 'Recursos' del curso (@ tecDigital), para que Ud. los descargue y estudie, a fin de modificarlos conforme a lo especificado en los proyectos del curso. En el repositorio también pueden encontrar un ambiente interactivo de edición, compilación y ejecución (IDE) desarrollado por el Dr. Luis Leopoldo Pérez (implementado en Java) y corregido por los estudiantes Pablo Navarro y Jimmy Fallas. Si desea aprender acerca de cómo integrar el IDE y el compilador, siga las indicaciones preparadas por nuestro ex-asistente, Diego Ramírez Rodríguez, en cuanto a las partes del compilador que debe desactivar para poder trabajar, así como los ajustes necesarios para que el IDE y el compilador funcionen bien conjuntamente. *No se darán puntos extra a los estudiantes que desarrollen su propio IDE; no* es objetivo de este curso *desarrollar* IDEs para lenguajes de programación. Sin embargo, si uno o más grupos de estudiantes aportan un IDE interesante, podemos considerarlo para futuras ediciones de este curso.

Estudie el capítulo 5 y los apéndices B y D del libro de Watt y Brown, así como el código del compilador, para comprender los principios y técnicas por aplicar, el lenguaje fuente original (Δ) y las interdependencias entre las partes del compilador. En clases dimos sugerencias sobre cómo abordar algunos de los retos que plantea la extensión de Δ sujeto de este proyecto (Δ_{xt}). Ud. extenderá el compilador que resultó de su trabajo en el Proyecto #1 de este curso (aunque es lícito basarse en trabajo de otros compañeros, según indicamos abajo).

Si su Proyecto #1 fue deficiente, *puede utilizar el programa desarrollado (para el Proyecto #1) por otr@s compañer@s como base para esta tarea programada, pero deberá pedir la autorización de reutilización a sus compañer@s y darles créditos explícitamente* en la documentación del proyecto que presenta el equipo del cual Ud. es miembro².

¹ El profesor es un ser humano, falible como cualquiera.

² Asegúrese de comprender bien la representación que sus compañeros hicieron para los árboles de sintaxis abstracta, la forma en que estos deben ser recorridos, y las implicaciones que tienen las decisiones de diseño tomadas por ellos, pero en el contexto de los requerimientos de este Proyecto #2.

Entradas

Los programas de entrada serán suministrados en archivos de texto. Los archivos fuente deben tener la extensión `.tri`. Si su equipo ‘domestica’ un IDE, como el suministrado por el profesor o algún IDE alternativo, puede usarlo en este proyecto. En tal IDE, el usuario debe ser capaz de seleccionar el archivo que contiene el texto del programa fuente por procesar, editarlo, guardarlo de manera persistente, compilarlo y enviarlo a ejecución (una vez compilado); además, deberá tener ‘pestañas’ semejantes a las que ofrece el IDE suministrado por el profesor.

Lenguaje fuente

Sintaxis

Remítase a la especificación del Proyecto #1 (`IC-5701 2021-2 Proyecto 1 v3.pdf`).

Contexto: identificación y tipos

Usaremos las reglas sintácticas del lenguaje Δ xt para indicar las restricciones de contexto. Interprete las reglas sintácticas desde una perspectiva *contextual*, esto es, vea en ellas *árboles de sintaxis abstracta* y no secuencias de símbolos. En lo que sigue usaremos las siguientes *meta-variables* para hacer referencia a árboles de sintaxis abstractas de las clases sintácticas indicadas³:

<ul style="list-style-type: none">• V: V-name• Id, N, Id_1, Id_2: Identifier• $Exp, Exp_1, Exp_2, Exp_3, Exp_i$: Expression• Com, Com_1, Com_2, Com_i: Command• Dec, Dec_1, Dec_2, Dec_3: Declaration• TD: Type-denoter• FPS: Formal-Parameter-Sequence• APS: Actual-Parameter-Sequence	<ul style="list-style-type: none">• PF_1, PF_2, PF_i, PF_n: Proc-Func• PFs: Proc-Func*• $Cases$: Cases• C: Case• Cs: Case*• Cl_1, Cl_2: Case-Literal• InL: Integer-Literal• ChL: Character-Literal
--	---

Expresiones

No hay cambios en las expresiones.

Comandos

El comando **skip** no requiere de revisión contextual, pues no depende de tipos ni de declaraciones de identificadores.

Para el *comando* de asignación, $V := Exp$ aplican las restricciones contextuales de Δ .

Considere estos comandos repetitivos añadidos a Δ :

```
repeat while  $Exp$  do  $Com$  end
repeat until  $Exp$  do  $Com$  end
repeat do  $Com$  while  $Exp$  end
repeat do  $Com$  until  $Exp$  end
```

Las restricciones contextuales son:

- Exp debe ser de tipo `Boolean`.
- Com y sus partes deben satisfacer las restricciones contextuales⁴.

En el comando de repetición controlada por contador

```
repeat for  $Id := range Exp_1 .. Exp_2$  do  $Com$  end
```

las restricciones son:

³ Usamos los subíndices de manera liberal.

⁴ Observe que ésta, como muchas de las restricciones contextuales, se formulan de manera recursiva. Es decir, ‘están bien’ en todos sus niveles de anidamiento.

- Exp_1 y Exp_2 deben ser ambas de tipo *entero*. Los tipos de Exp_1 y Exp_2 deben ser determinados en el contexto en el que aparece *este* comando **repeat_for_range_do_end**.
- Id es conocida como la “variable de control”. Id es de tipo *entero*. Id es *declarada* en este comando y su alcance es Com ; *esta* declaración de Id *no* es conocida por Exp_1 *ni* por Exp_2 .
- Com debe cumplir con las restricciones contextuales. Su contexto es el mismo que rige el comando **repeat_for_range_do_end** dentro del cual aparece Com , *extendido* por la declaración de la variable de control Id .
- La variable de control Id *no* puede aparecer a la izquierda de una asignación ni pasarse como parámetro **var**⁵ en la invocación de un procedimiento o función dentro del cuerpo del comando repetitivo **repeat_for_range_do_end**⁶.
- Este comando repetitivo *funciona como un bloque* al declarar una variable local (Id , la variable de control). Las reglas usuales de anidamiento aplican aquí (si se re-declara la variable en un comando anidado, en una expresión, en una lista de parámetros o en cualquier bloque anidado, esta es *distinta* y hace *inaccesible* la variable de control declarada en el comando **repeat_for_range_do_end**, que sería más externa).

En las variantes condicionadas del **repeat_for_range_do_end**,

```
repeat for  $Id$  := range  $Exp_1$  ..  $Exp_2$  while  $Exp_3$  do  $Com$  end
repeat for  $Id$  := range  $Exp_1$  ..  $Exp_2$  until  $Exp_3$  do  $Com$  end
```

aplican las restricciones anteriores y se extienden así:

- Exp_3 debe ser de tipo *booleano*. El tipo de Exp_3 debe ser determinado en el contexto en el que aparece *este* comando **repeat_for_range_do_end**, *extendido* por la variable de control Id .

En el comando de repetición para acceder a elementos de un arreglo

```
repeat for  $Id$  in  $Exp$  do  $Com$  end
```

deben satisfacerse estas restricciones:

- Exp debe ser de tipo **array** *InL of TD*.
- Id es conocida como la “variable de iteración”. Id es de tipo *TD*. Id es *declarada* en este comando y su alcance es Com ; *esta* declaración de Id *no* es conocida por Exp .
- Com debe cumplir con las restricciones contextuales. Su contexto es el mismo que rige el comando **repeat_for_in_do_end** dentro del cual aparece Com , *extendido* por la declaración de la variable de iteración Id .
- La variable de iteración Id *no* puede aparecer a la izquierda de una asignación ni pasarse como parámetro **var**⁷ en la invocación de un procedimiento o función dentro del cuerpo del comando repetitivo **repeat_for_in_do_end**⁸.
- Este comando repetitivo *funciona como un bloque* al declarar una variable local (Id , la variable de iteración). Las reglas usuales de anidamiento aplican aquí (si se re-declara la variable en un comando anidado, en una expresión, en una lista de parámetros o en cualquier bloque anidado, esta es *distinta* y hace *inaccesible* la variable de iteración declarada en el comando **repeat_for_in_do_end**, que sería más externa).

El comando condicional de Δx fue modificado solamente en cuanto a sintaxis. Considere:

```
if  $Exp$  then  $Com_1$  ( |  $Exp_i$  then  $Com_i$ )* else  $Com_2$  end
```

Las expresiones Exp y Exp_i deben ser de tipo *booleano*. Com_1 , Com_i y Com_2 , así como sus partes, deben satisfacer las restricciones contextuales.

⁵ No podrá pasarse por referencia.

⁶ En clase discutimos que, en el cuerpo de este comando repetitivo (Com), la ‘variable’ de control Id en realidad debe comportarse como una *constante*.

⁷ No podrá pasarse por referencia.

⁸ En clase discutimos que, en el cuerpo de este comando repetitivo (Com), la ‘variable’ de control Id en realidad debe comportarse como una *constante*.

El comando

let Dec in Com end

se analiza contextualmente de manera idéntica al comando correspondiente en Δ original (el original no tiene **end** y ahora permitimos **Command** en lugar del **single-Command** original, pero esos son asuntos sintácticos).

Considere el comando de selección **select Exp from Cases end**

- La expresión selectora *Exp* debe ser de tipo entero (*Integer*) o de tipo carácter (*Char*).
- Todas las literales de un mismo **select** deben ser del mismo tipo que el de la expresión selectora⁹.
- Cada literal entera *InL* o literal carácter *ChL* debe aparecer *una sola vez* como caso dentro de un *mismo select*¹⁰. Esto incluye a los sub-rangos: las intersecciones entre literales solitarias y rangos de literales deben ser vacías (en un mismo **select**).
- Los **selects** pueden anidarse y sus literales tienen *alcance*; una literal de un **select** externo puede *reaparecer* dentro de un **select** anidado y esto *no* es un error.
- Todos los comandos subordinados (que aparecen dentro del **select**) deben satisfacer las restricciones contextuales¹¹.

Declaraciones

Los parámetros de una abstracción (*función* o *procedimiento*) forman parte de un nivel léxico (alcance) inmediatamente más profundo que aquel en el cual aparece el identificador de la abstracción que los declara¹². *Usted debe verificar que ningún nombre de parámetro se repita dentro de la declaración de una función o procedimiento (en el encabezado).*

Veamos la declaración de variable inicializada:

var Id := Exp

Aquí, el tipo de la *variable* inicializada *Id* debe inferirse a partir del tipo de la expresión inicializadora *Exp*¹³. *Id* se "exporta" al resto del bloque donde aparece esta declaración.

En una declaración **local Dec₁ in Dec₂ end**, los identificadores declarados en *Dec₁* son conocidos *exclusivamente* en *Dec₂*. Únicamente se "exportan" los identificadores declarados en *Dec₂*. Observe que tanto *Dec₁* como *Dec₂* son declaraciones generales (*Declaration*)¹⁴.

En una declaración **recursive PFs end** se permite combinar las declaraciones de varios procedimientos o funciones, de manera que puedan invocarse unos a otros (lo que pretende posibilitar la recursión mutua).

- Consideremos primero el caso de *dos* declaraciones: **recursive PF₁ | PF₂ end**. Allí se declaran funciones y(o) procedimientos mutuamente recursivos: el identificador¹⁵ declarado en *PF₁* es conocido por *PF₂*, y viceversa.
- Los nombres de función o procedimiento que aparecen en *PF₁* y *PF₂* deben ser *distintos*.
- Las secuencias de parámetros que aparecen en *PF₁* y *PF₂* están en un nivel de profundidad léxica mayor que el de los nombres de la función o procedimiento que les corresponden.

⁹ Si la expresión selectora es de tipo entero, entonces las literales que aparecen en los casos de ese **select** deben ser literales enteras. Si la expresión selectora es de tipo carácter, entonces las literales que aparecen en los casos **select** deben ser caracteres (literales de un carácter).

¹⁰ Tengan particular cuidado con los valores comprendidos en un sub-rango $Cl_1 \dots Cl_2$. Ese sub-rango comprende todos los valores que están en el conjunto $\{ lit : Literal \mid Cl_1 \leq lit \leq Cl_2 \}$. *Literal* puede corresponder a *Integer* o a *Char*.

¹¹ Es decir, las restricciones contextuales deben cumplirse *recursivamente* (en profundidad) en los comandos, expresiones y declaraciones subordinados.

¹² Es decir, los parámetros se comportan como identificadores declarados localmente en el bloque de la función o procedimiento. Repase lo expuesto en el libro de Watt y Brown, el código del compilador de Δ original en Java (método *OpenScope*), así como lo discutido en clases.

¹³ El procesamiento requerido es semejante al que se hace para la declaración **const Id ~ Exp**. La diferencia es que, en el caso que nos ocupa, se está declarando a *Id* como una *variable*, **no** como una *constante*. Un identificador declarado como *variable* sí puede ser destino de una asignación o ser pasado por referencia en una invocación a un procedimiento o a una función.

¹⁴ No son *single-Declaration* ni *compound-Declaration*. Revise lo explicado en clases.

¹⁵ Ese identificador da nombre a la función o al procedimiento.

- Se "exportan"¹⁶ los identificadores de función o de procedimiento declarados en PF_1 y PF_2 . Tanto en PF_1 como en PF_2 , los *parámetros (FPS)* declarados en un encabezado de función o de procedimiento son *privados* (es decir, *locales*) y, por lo tanto, no se "exportan".
- En el caso general, es posible declarar dos o más procedimientos o funciones como mutuamente recursivos. Sintácticamente: **recursive** $PF_1 \mid PF_2 \mid \dots \mid PF_n$ **end**. Todos los identificadores de *función* o de *procedimiento* introducidos por las declaraciones PF_i ($1 \leq i \leq n$) deben ser *distintos* y son conocidos al procesar los *cuerpos* de *todas* las funciones o procedimientos declarados en las PF_i .
- Es un error declarar más de una vez el mismo identificador¹⁷ en las declaraciones simples que componen una misma declaración compuesta **recursive**.
- Todos los identificadores de procedimiento o de función declarados vía **recursive** son "exportados"; esto es, son agregados al contexto subsiguiente y son conocidos después de la declaración compuesta **recursive**¹⁸.
- Funciones o procedimientos *distintos* declarados vía **recursive** pueden declarar *parámetros* con nombres idénticos, pero en listas de parámetros distintas. Eso *no* es problema.

Proceso y salidas

Ud. modificará el procesador de Δ extendido que preparó para el Proyecto #1, de manera que sea capaz de procesar las restricciones contextuales especificadas arriba.

- El analizador contextual debe realizar completamente el trabajo de identificación (manejo del alcance en la relación entre ocurrencias de definición y ocurrencias de aplicación de los identificadores) y realizar la comprobación de tipos sobre el lenguaje Δ extendido; debe reportar la posición de *cada uno* de los errores contextuales detectados (esto es, avisar de *todos* los errores contextuales encontrados en el proceso).
- Las técnicas por utilizar son las expuestas en clase y en el libro de Watt y Brown. Recuerde las reglas contextuales que el profesor expuso en clase.
- Nos interesa que el analizador contextual deje el árbol sintáctico 'decorado' apropiadamente para la fase de generación de código subsiguiente (Proyecto #3)¹⁹. Esto es particularmente delicado para el procesamiento de la variante **for** del comando **repeat** y las nuevas formas de declaración compuesta.

Como hemos indicado, ustedes deben basarse en los programas que les han sido dados como punto de partida. Su programación debe ser consistente con el estilo presente en el procesador en Java usado como base, y ser respetuosa de ese estilo. En el código fuente deben incluir comentarios que indiquen claramente los lugares donde ustedes han introducido modificaciones.

Deben dar crédito por escrito a cualquier fuente de información o de ayuda que hayan consultado.

Debe ser posible activar la ejecución del IDE de su compilador desde el Explorador de Windows haciendo clics sobre el ícono de su archivo .jar, o bien generar un .exe a partir de su .jar. Por favor indique claramente cuál es el archivo ejecutable del IDE, para que el profesor o nuestros asistentes puedan someter a pruebas su procesador sin dificultades. Si Ud. trabaja en Linux, Mac OS o alguna variante de Unix, por favor avise al profesor y a los asistentes cuanto antes.

Documentación

Debe documentar clara y concisamente los siguientes puntos²⁰:

¹⁶ Después de la declaración **recursive** los identificadores de procedimiento o de función serán visibles y podrán ser utilizados por declaraciones subsecuentes.

¹⁷ De función o de procedimiento.

¹⁸ Esto no sucede si la declaración **recursive** aparece dentro de una declaración **local** antes del **in**.

¹⁹ Esto comprende introducir información de tipos en los árboles de sintaxis abstracta (ASTs) correspondientes a expresiones, así como dejar "amarradas" las ocurrencias aplicadas de identificadores (poner referencias hacia el subárbol donde aparece la ocurrencia de definición correspondiente), vía la tabla de identificación [esto son los "punteros rojos" que hemos mencionado varias veces en clases. También interesa determinar si un identificador corresponde a una variable, etc.

²⁰ Nada en la documentación es opcional. La documentación tiene un peso importante en la calificación del proyecto.

Analizador contextual

- Describir la manera en que se comprueban los tipos para todas las variantes de **repeat ... end**.
- Describir la solución dada al manejo de alcance, del tipo y de la protección de la variable de control del **repeat for ... end** y sus variantes condicionadas (**while** y **until**).
- Describir la solución creada para resolver las restricciones contextuales del comando **select Exp from Cases end**.
- Describir el procesamiento de la declaración de variable inicializada (**var Id := Exp**).
- Descripción de su validación de la unicidad²¹ de los nombres de parámetros en las declaraciones de funciones o procedimientos.
- Describir la manera en que se procesa la declaración compuesta **local**. Interesa que la primera declaración (antes del **in**) introduzca identificadores que son conocidos *privadamente (localmente)* por la segunda declaración (después del **in**). Al finalizar el proceso, se "exporta" solamente lo introducido por la segunda declaración.
- Describir el procesamiento de la declaración compuesta **recursive**. Interesa la no-repetición de los identificadores *función* o de *procedimiento* (**Proc-Funcs**) declarados por esa declaración compuesta y que estos identificadores sean conocidos en los *cuerpos* de todas las declaraciones de funciones o procedimientos (**Proc-Funcs**) que aparecen en una misma declaración compuesta **recursive**.
- Nuevas rutinas de análisis contextual, así como cualquier modificación a las existentes.
- Lista de nuevos errores contextuales detectados.
- Plan de pruebas para validar el compilador. Debe separar las pruebas que validan la fase de análisis contextual, para no confundirlas con las pruebas dirigidas a validar los analizadores léxico y sintáctico. Debe incluir pruebas *positivas* (para confirmar funcionalidad con datos correctos) y pruebas *negativas* (para evidenciar la capacidad del compilador para detectar errores). Debe especificar lo siguiente para cada caso de prueba:
 - Objetivo del caso de prueba
 - Diseño del caso de prueba
 - Resultados esperados
 - Resultados observados
- En las pruebas del analizador contextual es importante comprobar que los componentes de análisis léxico y sintáctico siguen funcionando bien. Si hacen correcciones a ellos, deben documentar los cambios en un apéndice de la documentación de este Proyecto #2.
- Discusión y análisis de los resultados obtenidos. Conclusiones a partir de esto.
- Una reflexión sobre la experiencia de modificar fragmentos de un compilador/ambiente escrito por terceras personas.
- Descripción resumida de las tareas realizadas por cada miembro del grupo de trabajo.
- Indicar cómo debe compilarse su programa.
- Indicar cómo debe ejecutarse su programa.
- Archivos con el texto fuente de su compilador. El texto fuente debe incluir comentarios que indiquen con claridad los puntos en los cuales se han hecho modificaciones.
- Archivos con el código objeto del compilador. **El compilador debe estar en un formato ejecutable directamente desde el sistema operativo Windows²², o llegar a un acuerdo alternativo con el profesor.**
- Debe guardar su trabajo en una carpeta comprimida (formato **zip**) según se indica abajo²³. Esto debe incluir:
 - Documentación indicada arriba, con una portada donde aparezcan los nombres y números de carnet de los miembros del grupo. Los documentos descriptivos deben estar en formato .pdf.
 - Código fuente, organizado en carpetas.
 - Código objeto. Recuerde que el código objeto de su compilador (programa principal) debe estar en un formato directamente ejecutable en Windows (o una alternativa, negociada con el profesor).

²¹ I.e. no-repetición.

²² En principio, se permitirá entregar el trabajo en otro ambiente, pero debe avisar de previo al profesor y a los asistentes.

²³ **No use** formato **rar**, porque es rechazado por el sistema de correo-e del TEC.

- Programas (.tri) de entrada que han preparado para probar su analizador contextual.

Puntos extra

Los grupos de hasta tres miembros podrán obtener puntos extra por su procesamiento completo del comando **select**. Los grupos de *cuatro* miembros obligatoriamente deben procesar el comando **select** completo.

Entrega

Fecha límite: **lunes 2021.11.08**, antes de las 23:55. No se recibirán trabajos después de la fecha y la hora indicadas.

Los grupos pueden ser de *hasta 4* personas.

Debe enviar por correo-e el *enlace*²⁴ a un archivo comprimido almacenado en la nube con todos los elementos de su solución a estas direcciones: itrejos@itcr.ac.cr, susana.cob.3@gmail.com (Susana Cob García, asistente), a.tapia1908@gmail.com (Alejandro Tapia Álvarez, asistente).

El asunto (subject) debe ser:

"IC-5701 - Proyecto 2 - " *carnet* " + " *carnet* " + " *carnet* " + " *carnet* "

Si su mensaje no tiene el asunto en la forma correcta, su proyecto será castigado con -10 puntos; podría darse el caso de que su proyecto no sea revisado del todo (y sea calificado con 0) sin responsabilidad alguna del profesor o de los asistentes (caso de que su mensaje fuera obviado por no tener el asunto apropiado). Si su mensaje no es legible (por cualquier motivo), su carpeta está en formato **.rar**, o contiene un virus, la nota será **0**.

La documentación vale alrededor de un 25% de la nota de cada proyecto. La redacción y la ortografía deben ser correctas. El profesor tiene *altas expectativas* respecto de la calidad de los trabajos escritos y de la programación producidos por estudiantes universitarios de tercer año de la carrera de Ingeniería en Computación del Tecnológico de Costa Rica. Los profesores esperamos que los estudiantes tomen en serio la comunicación profesional.

²⁴ Los sistemas de correo del TEC rechazan el envío o la recepción de carpetas comprimidas con componentes ejecutables. Suban su carpeta comprimida (en formato **.zip**) a algún 'lugar' en la nube y envíen el hipervínculo al profesor y a nuestros asistentes mediante un mensaje de correo con el formato indicado. Asegúrense que el hipervínculo da derechos de *lectura* a profesor y asistentes.