Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
IC-5701 Compiladores e intérpretes

Proyecto #3, Las fases de síntesis: generación e interpretación de código

Historial de revisiones:

- 2019.10.31, 2019.11.05: Versión base. Comunicaciones orales sobre alcances, más ejemplos en clase.
- 2019.11.12: Versión v0. Color azul para problemas que dan puntos extra.

Lea con cuidado este documento. Si encuentra errores en el planteamiento[1], por favor comuníqueselos inmediatamente al profesor.

Objetivo

Al concluir este tercer proyecto, Ud. habrá terminado de comprender los detalles relativos a las fases de síntesis de un procesador del lenguaje Δ extendido (generación de código para la máquina virtual TAM e interpretación de ese código). Al igual que en los proyectos anteriores, el compilador y el intérprete aplican los principios, los métodos y las técnicas expuestos por Watt y Brown en su libro *Programming Language Processors in Java*. Ud. deberá extender el par (compilador, intérprete) de (Δ, TAM) desarrollado por Watt y Brown, de manera que sea capaz de procesar el lenguaje Δ extendido, según se describe en la sección *Lenguaje fuente* de la especificación de los proyectos #1 y #2, y conforme con las indicaciones que se hacen abajo. Además, su compilador deberá coexistir con el ambiente de edición, compilación y ejecución ("IDE"), construido en Java por el Dr. Luis Leopoldo Pérez y ajustado por estudiantes de Ingeniería en Computación del TEC.

En negro se plantean los temas que deberá cubrir en este proyecto ***obligatoriamente***.

En azul se plantean los temas que darán puntos extra.

Base

La base es la misma dada para los proyectos #1 y #2. Debe estudiar y comprender los capítulos y apéndices del libro *Programming Language Processors in Java* correspondientes a *organización en tiempo de ejecución*, *generación de código*, *interpretación* y *descripción de TAM* (6, 7, 8, B, C, D). Deberá estudiar y entender la estructura y las técnicas empleadas en la construcción del generador de código que traduce programas de Triángulo extendido (`.tri`) hacia código de la Máquina abstracta TAM; estos componentes están en la carpeta 'CodeGenerator'. También deberá estudiar el intérprete de TAM, cuyos componentes están en la carpeta 'TAM'.

Si su proyecto ***anterior*** fue deficiente, *puede utilizar el programa desarrollado por otros compañeros como base para esta tarea programada*, ***pero deberá pedir la autorización de reutilización de sus compañeros y darles créditos explícitamente*** en la documentación del proyecto que presenta el equipo del cual Ud. es miembro[2].

Entradas

Los programas de entrada serán suministrados en archivos de texto corrientes (ASCII). El usuario seleccionará cuál es el archivo del programa fuente desde el ambiente de programación (IDE) base suministrado, o bien lo editará en la ventana que el IDE provee para el efecto (que permitirá guardarlo de manera persistente). Los archivos fuente deben tener terminación `.tri`.

Lenguaje fuente

Sintaxis

Remítase a la especificación del Proyecto #1 ('IC-5701 2019-2 Proyecto 1'). ***Asegúrese de presentar los árboles de sintaxis abstracta en la ventana correspondiente***, en caso de que no lo hubiera hecho en los proyectos anteriores.

---

[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.

Contexto: identificación y tipos
No hay cambios.

Es importante que entienda bien cómo funciona la variable de control en el comando de repetición controlada por contador, **`loop for`** *Id* **`~`** *Exp*$_1$ **`to`** *Exp*$_2$ **`do`** *Com* **`repeat`**, para que le dé la semántica deseada (alcance léxico, efecto en la estructura de bloques, variable protegida para que no pueda ser modificada vía asignación o paso de parámetros por referencia). Asegúrese de que el analizador contextual haya dejado el árbol sintáctico decorado apropiadamente, esto es, que haya introducido información de tipos en los árboles sintácticos correspondientes a expresiones, así como dejar "amarradas" las ocurrencias *aplicadas* de identificadores vía la tabla de identificación (poner referencias hacia el subárbol sintáctico donde aparece la ocurrencia de *definición* correspondiente, es decir, la declaración de la variable *Id* asociada a un valor inicial dado por la expresión *Exp*$_1$). Asimismo, deberá determinar si un identificador corresponde a una variable[3], etc. Refiérase a la especificación del proyecto #2 y repase el capítulo 5 del libro.

Semántica
Esta es la parte que influye en la generación e interpretación de código.

Solo se describe la semántica de las frases de Δ que han pasado el análisis contextual. No se debe generar código para programas con errores contextuales, sintácticos o léxicos.

Para los valores declarados como de tipo **`array`** IL **`of`** T, el primer elemento tiene índice 0 y el último elemento tiene índice IL-1; hay IL elementos en arreglos definidos de esta manera. En un acceso a arreglo, el valor que resulta de evaluar la expresión seleccionadora (indexadora) debe estar dentro del rango de los índices; de lo contrario se deberá ***abortar*** la ejecución del programa con un mensaje de error apropiado, como "`array out of bounds`". El compilador original no genera código que compruebe índices dentro de rango al acceder arreglos.

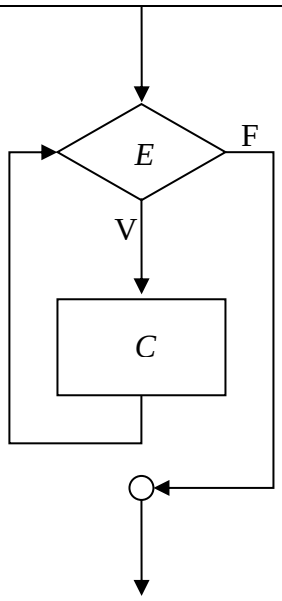
El comando nulo **`skip`** no tiene efectos en la ejecución: no se afectan la memoria, ni la entrada ni la salida. Tenga cuidado con el esquema de generación de código para que de veras el comportamiento sea "nulo"[4].

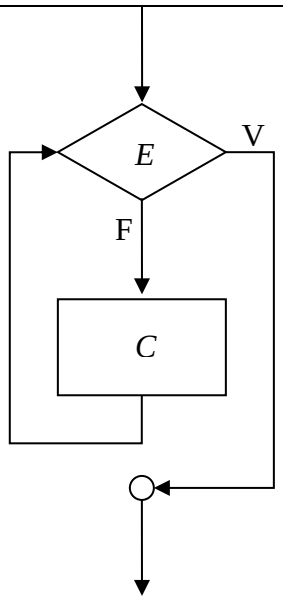
En el comando **`if`** *Exp* **`then`** *Com*$_1$ **`else`** *Com2* **`end`** se procede como en el lenguaje Δ original.

Los comandos repetitivos pueden ser ejecutados cero, una o más veces, según se indica a continuación.

`loop while` *E* **`do`** *C* **`repeat`**	**`loop until`** *E* **`do`** *C* **`repeat`**
(*Repetición con entrada condicionada positiva*) Se evalúa la expresión *E*. Si ésta es verdadera, se procede a ejecutar el comando *C*; luego se procede a re-evaluar *E* y determinar si se repite *C* o no. Si *E* evalúa a falso, se termina la repetición. Note que *C* podría ejecutarse 0 veces (si *E* evalúa a falso al inicio).	(*Repetición con entrada condicionada negativa*) Se evalúa la expresión *E*. Si ésta es falsa, se procede a ejecutar el comando *C*; luego se procede a re-evaluar *E* y determinar si se repite *C* o no. Si *E* evalúa a verdadero, se termina la repetición. Note que *C* podría ejecutarse 0 veces (si *E* evalúa a verdadero al inicio).

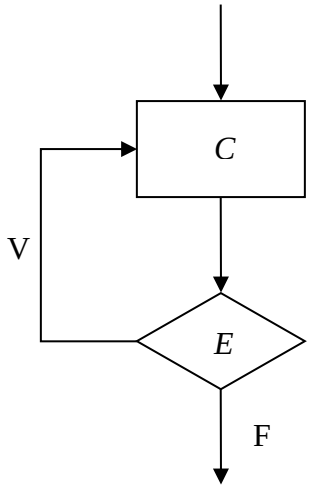
[3] Como discutimos ampliamente en clases, la variable de control de un **`loop_for_~_to_do_end`** se comporta como una constante (no se la puede asignar, no se la puede pasar por referencia), pero es preferible usar un constructor propio para esta situación: *no el de una constante*.
[4] Pista: el comando nulo ya existía en el compilador original. Estudie eso.





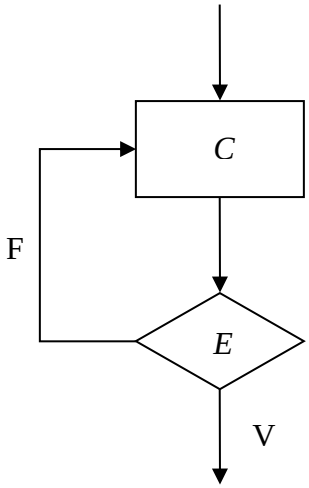
`loop do` *C* `while` *E* `repeat`

(*Repetición con entrada garantizada y salida negativa*) Se ejecuta el comando *C*. Se evalúa la expresión *E*. Si *E* evalúa a verdadera, se procede a ejecutar el comando *C* de nuevo. Si *E* evalúa a falsa, se termina la repetición. Note que *C* se ejecuta al menos una vez.



`loop do` *C* `until` *E* `repeat`

(*Repetición con entrada garantizada y salida positiva*) Se ejecuta el comando *C*. Se evalúa la expresión *E*. Si *E* evalúa a falsa, se procede a ejecutar el comando *C* de nuevo. Si *E* evalúa a verdadera, se termina la repetición. Note que *C* se ejecuta al menos una vez.



El comando

loop for *Id* **~** *Exp*$_1$ **to** *Exp*$_2$ **do** *Com* **repeat**

declara una variable *Id* que es ***local*** a la repetición[5] y cumple con las restricciones indicadas en el Proyecto #2. Tal comando se ***comporta*** como el código que sigue[6]:

```
let const $Final ~ Exp2 ; !el valor final se evalúa solo una vez
    var   Id init Exp1 !Id obtiene como primer valor el que tiene Exp1
in loop while Id <= $Final do !mientras no se haya excedido el límite superior
     let
       const Id ~ Id !se re-declara Id como constante para usarlo en Com
                     !esto protege a Id dentro de Com, dada la estructura de bloques de Δ
     in Com
     end ; !el let interno abarca únicamente Com, que llega hasta aquí
     Id := Id + 1 !se incrementa la variable de control, Id, declarada en el primer let
     !continuar con las repeticiones
   repeat
end
```

En particular observe que:

- La expresión *Exp*$_2$ se evalúa *una sola vez*, al inicio de la repetición (por eso se asocia el valor a una constante, `$Final`), y debe ser entera – como lo garantiza el analizador contextual. Si el valor se guarda en memoria, ese espacio debe ser liberado al terminar la repetición.
- La expresión *Exp*$_1$ se evalúa *una sola vez*, al inicio de la repetición, debe ser entera y ése es el valor *inicial* de la variable de control `Id`.
- La variable de control (`Id`) es *declarada* por el comando **`loop_for_~_to_end`**; cuando la repetición termina, esta variable debe desaparecer de la memoria[7].
- La variable de control *debe 'funcionar' como una constante* en el cuerpo del **`loop_for_~_to_end`**(`Com`): en el comando `Com` no se le pueden asignar valores; estas restricciones se suponen aseguradas por el analizador contextual, al 'instalar' adecuadamente una asociación en el ambiente. El generador de código debe emitir las instrucciones necesarias para *actualizar* el valor de la variable de control (`Id`)[8].

Considere la declaración de variable inicializada:

var *Id* **init** *Exp*

En esta declaración se evalúa la expresión, cuyo valor resultante queda en la *cima* de la pila de TAM[9] y da valor inicial a la variable. La dirección donde inicia ese valor debe asociarse al identificador (*Id*) de la variable; recuerde que las direcciones se forman mediante desplazamientos relativos al contenido de un registro que sirve como *base*. *Exp* puede ser de cualquier tipo; el almacenamiento requerido para guardar la variable *Id* dependerá del tamaño de los datos del *tipo* inferido para *Exp* (lo cual ya fue hecho por el analizador contextual).

---

[5] La 'variable de control' se crea para cada activación del **`loop_for_~_to_end`**. El espacio que esta ocupe debe ser liberado al terminar la ejecución de dicho comando iterativo, así como cualquier otro espacio (en memoria) requerido para ejecutar ese comando.

[6] Ese código es para *explicar* el comportamiento del **`loop_for_~_to_end`**. Que se comporte así *no implica* que Ud. deba transliterar esta descripción y hacer una generación de código rudimentaria basada en ella. La pseudo-constante `$Final` ha sido introducida únicamente para *explicar* el comportamiento de este comando repetitivo. Tal constante no es parte del código del programa fuente. Haga una interpretación clara y eficiente de la semántica deseada para este comando.

[7] Es decir, Id sale de la pila (se desaloja de memoria).

[8] En la explicación, la *constante* `Id` en el **`let`** interno toma el valor actual de la *variable* `Id` del **`let`** externo. La declaración de `Id` en el **`let`** interno asegura que `Com` no puede usar a `Id` como variable. *Después* de ejecutar el **`let`** interno se actualiza la variable de control. Pero esto solo *explica* el funcionamiento de este comando repetitivo; una vez comprendida la semántica del comando, Ud. debe desarrollar un buen esquema de generación de código para el comando – en clases ofrecimos una plantilla que puede ayudarle a plantear su propio esquema de generación de código para el comando **`loop_for_~_to_end`**. Recuerde que el analizador contextual debió asegurar que `Id` no puede ser modificada vía asignaciones por *Com* ni puede ser pasada por referencia.

[9] Debe quedar inmediatamente debajo de la celda de memoria apuntada por el registro `ST`, cuando este haya sido actualizado.

Debe asignarse espacio y direcciones aparte para *cada una* de las entidades declaradas dentro de las declaraciones compuestas **`local`** y **`recursive`**[10]. Las entidades declaradas en un **`recursive`** deben conocer las direcciones de todas las demás entidades introducidas en la misma declaración[11]. Recuerde que **`local`** exporta entidades, pero las entidades exportadas deben tener acceso a lo declarado como *local* ('privado'). Desde la perspectiva de generación de código, el procesamiento de las declaraciones compuestas **`local`** y **`recursive`** no ofrece otras complicaciones adicionales a las que presenta el procesamiento de las declaraciones secuenciales. Suponemos que el analizador contextual dejó correctamente establecidas las referencias dentro del AST decorado (los "punteros rojos").

El resto de las características deberán ser procesadas como para el lenguaje Δ original.

Generación de código y modificaciones a TAM

Podríamos modificar el repertorio de instrucciones de TAM, en cuanto a acceder arreglos de Δ, para comprobar, *en tiempo de ejecución*, que el valor correspondiente a la expresión seleccionadora (que indiza) está dentro del rango: entre 0 e `IL` – 1 (ambos inclusive) para arreglos declarados de tipo **`array`** `IL` **`of`** `T`. Una posibilidad es introducir una nueva ***primitiva***, *`indexcheck`*, que acceda los tres elementos superiores de la pila, los cuales serían[12]: cota superior, cota inferior e índice. Cuando el índice está fuera de rango debe abortarse la ejecución, indicando la naturaleza del error. Una alternativa es que la plantilla de generación de código para el acceso a arreglos tenga instrucciones `JUMPIF` idóneas para hacer las comparaciones con las cotas del (tipo) del arreglo[13].

Proceso y salidas

Ud. modificará el procesador original de Δ y el intérprete de TAM, ambos en Java, para que sean capaces de procesar las extensiones especificadas arriba.

- Las técnicas por utilizar son las expuestas en clase y en el libro de Watt y Brown.
- Debe documentar *todas* las plantillas de generación de código correspondientes a las extensiones de Δ implementadas por su procesador.
- El algoritmo de generación de código debe corresponder a sus plantillas de generación de código.
- Su programación debe ser consistente con el estilo aplicado en los programas en Java usados como base (compilador de Δ, intérprete de TAM), y ser respetuosa de ese estilo. En el código fuente debe estar claro dónde ha introducido Ud. sus modificaciones.
- Las salidas de la ejecución del programa Triangle, así como las entradas al programa, deben aparecer en la pestaña 'Console' del IDE.
- Los árboles de sintaxis abstracta deben desplegarse en la pestaña 'Abstract Syntax Trees' del IDE. Esto se logra extendiendo el trabajo que ya hace el IDE sobre el lenguaje Triangle original: visitando los ASTs y construyendo subárboles apropiados que se presentan gráficamente en la pestaña.
- El código generado debe ser escrito en un archivo que tiene el mismo nombre del programa fuente, pero con extensión `.tam`. El archivo `.tam` debe quedar en la misma carpeta donde está el código fuente.
- El código generado debe aparecer en la pestaña 'TAM Code' del IDE. Esto se logra extendiendo el trabajo que ya hace el IDE al desensamblar el código TAM. Considere los cambios realizados a TAM, para que la salida sea significativa.
- Las entidades declaradas deberán aparecer en la pestaña 'Table Details' del IDE. En particular, nos interesa que el generador de código añada información apropiada en el árbol de sintaxis abstracta para poder reportar claramente las entidades declaradas (ya que esto se obtiene al recorrer el árbol de sintaxis abstracta (decorado) para obtener los descriptores de entidades en tiempo de ejecución, y *no* de una tabla de identificación).

---

[10] Recuerde que las entidades que pueden declararse mediante **`recursive`** son procedimientos y funciones, por lo que el espacio y las direcciones son en memoria de *código*.

[11] En las declaraciones **`recursive`**, las direcciones son de código, pues lo que se declara como mutuamente recursivo son procedimientos y funciones. El procesamiento de **`recursive`** requiere dos pasadas sobre el árbol de esa declaración compuesta, como se sugirió en clases, para poder resolver las referencias 'hacia adelante' mediante un proceso de "parchado" en que puede aprovecharse un patrón 'publicador-suscriptores'. Por lo demás, desde la perspectiva de la generación de código, el procesamiento de las declaraciones compuestas **`recursive`** se asemeja al procesamiento de las declaraciones secuenciales de procedimientos y funciones.

[12] Desde la cima (el "top") hacia "abajo", es decir, desde el `ST` hacia el `SB`.

[13] ¡Cuidado! Recuerde que `JUMPIF` saca valores de la pila una vez realizada la comparación.

- Si un programa terminase anormalmente (aborta), en la consola debe indicarse claramente la razón. Escríbala en inglés, para ser consistente con los demás mensajes de error.
- El código TAM solo puede ser ejecutable cuando la compilación del programa fuente en Δ extendido está libre de errores léxicos, sintácticos y contextuales.
- **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 su asistente puedan someter a pruebas su programa sin dificultades*.
- *Debe dar crédito por escrito a cualquier fuente de información o ayuda.*

Documentación

Debe documentar clara y concisamente los siguientes puntos[14]:

- Descripción del esquema de generación de código para el comando **`skip`**.
- Descripción del esquema de generación de código para el comando **`if_then_else_end`**.
- Descripción del esquema de generación de código para *todas* las variantes de **`loop`** … **`end`**.
- Descripción del esquema de generación de código para **`loop for`** … **`end`**.
- Solución dada al procesamiento de declaraciones de variable inicializada (**`var`** `Id` **`init`** `Exp`).
- Su solución al problema de introducir declaraciones privadas (**`local`**).
- Su solución al problema de introducir declaraciones de procedimientos o funciones mutuamente recursivos (**`recursive`**).
- Nuevas rutinas de generación o interpretación de código, así como cualquier modificación a las existentes.
- Extensión realizada a los procedimientos o métodos que permiten visualizar la tabla de nombres (aparecen en la pestaña 'Table Details' del IDE).
- Descripción de la manera en que se resuelve la comprobación de cotas en el acceso a arreglos (en tiempo de ejecución).
- Descripción de cualquier modificación hecha a TAM y a su intérprete.
- Lista de errores de generación de código o de ejecución detectados.
- Describir cualquier cambio que requirió hacer al analizador sintáctico o al contextual para efectos de generación de código (incluida la representación de árboles sintácticos).
- Describir cualquier modificación hecha al ambiente de programación para hacer más fácil de usar su compilador.
- Dar crédito a los autores de cualquier programa utilizado como base (esto incluye el IDE y el código de los proyectos #1 y #2).
- Plan de pruebas para validar el compilador. Debe separar las pruebas para el generador de código de las que validan al intérprete. 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
- Discusión y análisis de los resultados observados. Conclusiones obtenidas a partir de esto.
- Descripción resumida de las tareas realizadas por cada miembro del grupo.
- Breve reflexión sobre la experiencia de modificar fragmentos de un (compilador | intérprete | ambiente) escrito por terceras personas, así como de trabajar en grupo para los proyectos de este curso.
- Indicar cómo debe compilarse su programa.
- Indicar cómo debe ejecutarse su programa.
- Una carpeta que contenga:
    - Carpetas donde se encuentre el texto fuente de sus programas. El texto fuente debe indicar con claridad los puntos en los cuales se han hecho modificaciones.

[14] *Nada* en la documentación es opcional. La documentación tiene un peso importante en la calificación del proyecto. Si no modifica algo que es requerido para este proyecto, indíquelo explícitamente.

- Carpetas donde se encuentre el código objeto del compilador+intérprete, en formato directamente ejecutable desde el sistema operativo Windows[15]. Todo el contenido del archivo comprimido debe venir libre de infecciones. Debe incluir el IDE enlazado a su compilador+intérprete, de manera que desde él se pueda ejecutar sus procesadores de Δ y de TAM.

- Debe reunir su trabajo en un archivo comprimido en formato **`.zip`**. 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.
  - Programas (`.tri`) de prueba que han preparado.

Entrega

Fecha límite: **lunes 2019.11.25** 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***[16] a un archivo comprimido almacenado en la nube con todos los elementos de su solución a estas direcciones: `itrejos@itcr.ac.cr` e `ignacio.gomez.chaverri@gmail.com` (Ignacio Gómez Chaverri, nuestro asistente). Asegúrese de dar los permisos de lectura y descarga suficientes para el profesor y su asistente.

> El asunto (subject) debe ser:
> "IC-5701 - Proyecto **3** - " <carnet> " + " <carnet> " + " <carnet> " + " <carnet>.
>
> Los carnets deben ir ordenados ascendentemente.

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 su asistente (caso de que su mensaje fuera obviado por no tener el asunto apropiado). Si su mensaje no es legible (por cualquier motivo), o contiene un virus, la nota será 0.

Estén atentos a recibir notificación respecto de posible entrega alternativa de trabajos vía el tecDigital.

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.

---

[15] Es decir, sin necesidad de compilar los programas fuente.

[16] Los sistemas de correo han estado rechazando 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 su asistente mediante un mensaje de correo con el formato indicado.