UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



ESCOLA TE´CNICA SUPERIOR DE ENXEN˜ AR´IA

Perldoop 2.0

Un compilador fuente-a-fuente Perl-Java

*Autor:*

César Piñeiro Pomar

*Directores:*

Juan Carlos Pichel Campos

José Manuel Abuín Mosquera

Grao en Enxeñeira Informática

Xulio 2016

Traballo de Fin de Grao presentado na Escola Técnica Superior de Enxeñería da universidade de Santiago de Compostela para a obtención do Grao en Enxeñaría Informática



**D. Juan Carlos Pichel Campos**, Profesor do Departamento de Electrónica e Computación da Universidade de Santiago de Compostela, e **D. José Manuel Abuín Mosquera**, Profesor do Departamento de Electrónica e Computación da Universidade de Santiago de Compostela,

INFORMAN:

Que a presenta memoria, titulada *Perldoop 2.0: Un compilador fuente-a-fuente Perl-Java*, presentada por **D. César Piñeiro Pomar** para superar os créditos correspondentes ao Traballo de Fin de Grao da titulación de Grao en Enxería Informática, realizouse baixo a nosa dirección no Departamento de Electrónica e Computación da universidade de Santiago de Compostela.

E para que así conste aos efectos oportunos , expiden o presente informe en Santiago de Compostela, a (Data):

O director, O codirector, O alumno,

Juan Carlos Pichel Campos José Manuel Abuín Mosquera César Piñeiro Pomar

Agradecementos

Se se quere pór algún agradecemento, este vai aquí

Índice

[1. Introducción 1](#_Toc454476984)

[1.1. Perldoop Original 1](#_Toc454476985)

[1.2. Problemas de la versión original 2](#_Toc454476986)

[1.3. Objetivos 3](#_Toc454476987)

[2. Gestión del Alcance 4](#_Toc454476988)

[2.1. Descripción del alcance 4](#_Toc454476989)

[2.2. Entregables 4](#_Toc454476990)

[2.3. Criterios de Aceptación 4](#_Toc454476991)

[2.4. Restricciones del Proyecto. 5](#_Toc454476992)

[2.5. Exclusiones del Proyecto 5](#_Toc454476993)

[2.6. Supuestos del Proyecto 5](#_Toc454476994)

[3. Casos de Uso 6](#_Toc454476995)

[4. Análisis de requisitos 12](#_Toc454476996)

[4.1. Requisitos funcionales 12](#_Toc454476997)

[4.2. Requisitos no funcionales 15](#_Toc454476998)

[5. Gestión del Proyecto 16](#_Toc454476999)

[5.1. Metodología 16](#_Toc454477000)

[5.1.1. Programación extrema 17](#_Toc454477001)

[5.2. Gestión de la configuración 17](#_Toc454477002)

[5.2.1. Identificación de elementos de configuración 18](#_Toc454477003)

[5.2.2. Herramienta gestión de la configuración 19](#_Toc454477004)

[5.3. Planificación temporal 19](#_Toc454477005)

[5.3.1. Planificación y Alcance 20](#_Toc454477006)

[5.3.2. Formación 21](#_Toc454477007)

[5.3.3. Análisis 21](#_Toc454477008)

[5.3.4. Diseño 21](#_Toc454477009)

[5.3.5. Implementación 22](#_Toc454477010)

[5.3.6. Pruebas 22](#_Toc454477011)

[5.3.7. Documentación 22](#_Toc454477012)

[5.3.8. Cronograma 23](#_Toc454477013)

[5.3.9. Diagrama de Gantt 24](#_Toc454477014)

[5.4. Gestión de riesgos 25](#_Toc454477015)

[5.4.1. Especificación 26](#_Toc454477016)

[5.5. Análisis de costes 29](#_Toc454477017)

[6. Herramientas y Tecnologías 31](#_Toc454477018)

[6.1. IDEs de desarrollo 31](#_Toc454477019)

[6.2. Tecnologías 31](#_Toc454477020)

[6.3. Aplicaciones 32](#_Toc454477021)

[7. Arquitectura del software 33](#_Toc454477022)

[7.1. Interfaz línea comandos 34](#_Toc454477023)

[7.2. Analizador Léxico 36](#_Toc454477024)

[7.3. Analizador Sintáctico 36](#_Toc454477025)

[7.4. Comprobación y generación de código 36](#_Toc454477026)

[7.5. Sistema de errores 37](#_Toc454477027)

[7.6. Sistema de mensajes 37](#_Toc454477028)

[8. Diseño e Implementación 38](#_Toc454477029)

[8.1. Diagrama de clases 38](#_Toc454477030)

[8.2. Diagramas de interacción 42](#_Toc454477031)

[8.3. Diseño tokens analizador léxico 44](#_Toc454477032)

[8.3.1. Tokens Básicos 44](#_Toc454477033)

[8.3.2. Tokens de palabras reservadas 46](#_Toc454477034)

[8.3.3. Tokens de Etiquetas 47](#_Toc454477035)

[8.4 Gramática Analizador Sintáctico 48](#_Toc454477036)

[8.4. Generación y Comprobación de código 56](#_Toc454477037)

[8.4.1. Tipos de datos 56](#_Toc454477038)

[8.4.1.1. Tipos asociados a etiquetas 56](#_Toc454477039)

[8.4.1.2. Tipos asociados a contantes 57](#_Toc454477040)

[8.4.2. Casting automático 57](#_Toc454477041)

[8.4.3. Definición de funciones 59](#_Toc454477042)

[8.4.4. Asignación 61](#_Toc454477043)

[8.4.5. Ficheros 63](#_Toc454477044)

[8.4.6. Paquetes 64](#_Toc454477045)

[8.4.7. Funciones nativas 65](#_Toc454477046)

[8.4.8. Expresiones regulares 66](#_Toc454477047)

[8.4.9. Hadoop 66](#_Toc454477048)

[8.4.9.1. Mapper 67](#_Toc454477049)

[8.4.9.2. Reducer 69](#_Toc454477050)

[8.5. Librería java 71](#_Toc454477051)

[8.5.1. Clases para Tipos 72](#_Toc454477052)

[8.5.2. Clases para Funciones 73](#_Toc454477053)

[9. Pruebas 74](#_Toc454477054)

[Bibliografía 75](#_Toc454477055)

Índice Figuras

[Figura 1: Plantilla Perldoop 2](#_Toc454477078)

[Figura 2: Diagrama casos de uso 6](#_Toc454477079)

[Figura 3: Estructura de descomposición de trabajo 20](#_Toc454477080)

[Figura 4: Diagrama Gantt parte 1 24](#_Toc454477081)

[Figura 5: Diagrama Gantt parte 2 25](#_Toc454477082)

[Figura 6: Arquitectura del sistema 34](#_Toc454477083)

[Figura 7: Ayuda Perldoop 2 35](#_Toc454477084)

[Figura 8: Ejemplo sistema de errores 37](#_Toc454477085)

[Figura 9: Clases con sus atributos y métodos 38](#_Toc454477086)

[Figura 10: Diagrama de clases 39](#_Toc454477087)

[Figura 11: Diagrama de secuencia de un análisis 42](#_Toc454477088)

[Figura 12: Error en el análisis 43](#_Toc454477089)

[Figura 13: Gramática analizador sintáctico 49](#_Toc454477090)

[Figura 14: Función perl con array y escalar 59](#_Toc454477091)

[Figura 15: Funcion Perl con referencia Array y escalar 59](#_Toc454477092)

[Figura 16: Ejemplo de tipado de funciones y su traducción 60](#_Toc454477093)

[Figura 17: Asignación entre listas 61](#_Toc454477094)

[Figura 18: Asignación entre lista y expresión 62](#_Toc454477095)

[Figura 19: Asignación entre lista y función 62](#_Toc454477096)

[Figura 20: Inicialización de colecciones 62](#_Toc454477097)

[Figura 21: Inicialización de colecciones optimizada 63](#_Toc454477098)

[Figura 22: Apertura de fichero 63](#_Toc454477099)

[Figura 23: Ejemplo de lectura de fichero 63](#_Toc454477100)

[Figura 24: Ejemplo de escritura de fichero 64](#_Toc454477101)

[Figura 25: Creación y uso de un paquete con su traducción 64](#_Toc454477102)

[Figura 26: Expresión regular match 66](#_Toc454477103)

[Figura 27: Mapper Perl 68](#_Toc454477104)

[Figura 28: Mapper java 68](#_Toc454477105)

[Figura 29:Reducer Perl 70](#_Toc454477106)

[Figura 30: Reducer java 71](#_Toc454477107)

[Figura 31: Clases para tipos 72](#_Toc454477108)

Índice Tablas

[Tabla 1: Tipos de documento 18](#_Toc454477109)

[Tabla 2: Cronograma 23](#_Toc454477110)

[Tabla 3: Tokens básicos 46](#_Toc454477111)

[Tabla 4: Tokens de palabras reservadas 47](#_Toc454477112)

[Tabla 5: Exprsiones regualres para etiquetas 47](#_Toc454477113)

[Tabla 6: Tokens de etiquetas 48](#_Toc454477114)

[Tabla 7: Relación etiqueta tipo de dato 56](#_Toc454477115)

[Tabla 8: Tabla de funciones auxiliares librería 73](#_Toc454477116)

# Introducción

Perldoop 2.0 es una herramienta que permite convertir código fuente escrito en Perl a código fuente escrito en java. Esta herramienta es una reconstrucción desde cero del Perldoop original con el objetivo de eliminar limitaciones y añadir nuevas características.

## Perldoop Original

El trabajo realizado por los investigadores del CiTIUS dio como resultado la creación del programa Perldoop[1]. Su diseño permite transformar de forma sencilla y eficiente un script Perl[2] para El Procesamiento de Lenguaje Natural (PLN) en un programa java compatible con las tecnologías Big Data (en concreto, Hadoop[3]).

La herramienta hace uso de etiquetas, introducidas en el código original, para poder convertir automáticamente un código Perl secuencial en un código Java adaptado al paradigma MapReduce mejorando su ejecución en un clúster.

Perldoop está basado en la ejecución de expresiones regulares sobre el código para aplicar las transformaciones, por esta razón la herramienta requiere una estructura bien definida para poder funcionar, en particular se necesitan dos ficheros para generar el código java.

1. En primer lugar, como es de esperar, un archivo que contenga el script perl a traducir. Este código debería:
   1. Seguir unas reglas estrictas de programación [4].
   2. Estar etiquetado correctamente [5].
2. Una platilla java es necesaria para asegurarse del correcto funcionamiento del código generado por esta herramienta. La plantilla debería incluir el constructor de la clase, así como los import de las dependencias necesarias.

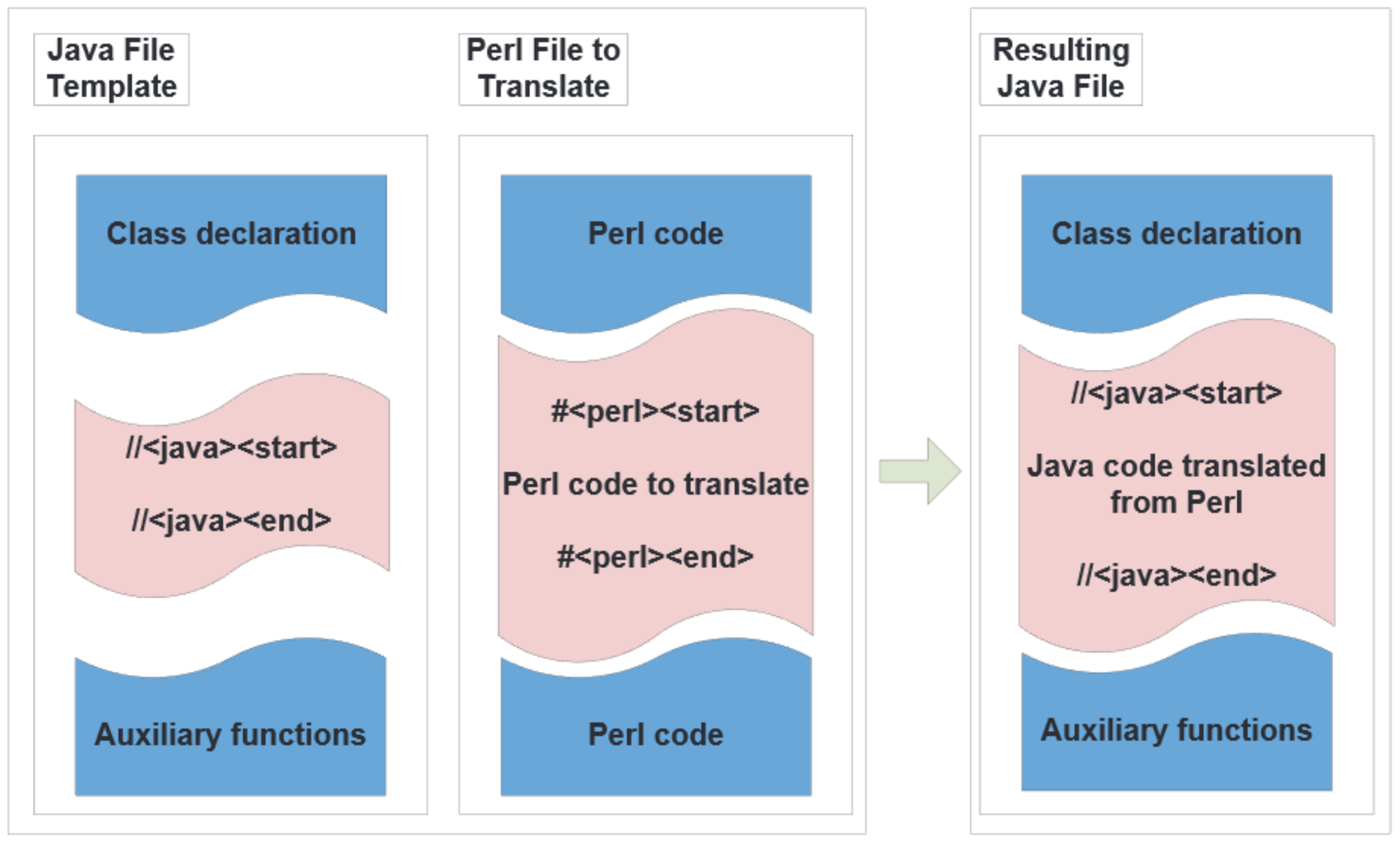


Figura 1: Plantilla Perldoop

La figurar 1.1 muestra el funcionamiento de perldoop, la herramienta traduce el código entre las etiquetas <perl><start> y <perl><end> del código fuente perl y pega esta traducción entre las etiquetas <java><start> y <java><end> de la plantilla.

## Problemas de la versión original

El principal problema de esta versión era la necesidad de una plantilla hecha a mano para poder ejecutar el código traducido. Una persona sin conocimientos de java necesitaba que alguien elaborase la plantilla y en caso de cambios en el script, dependencias sobre todo, podía ser necesario tener que cambiar la plantilla. La fuente del problema es que, al usar patrones, solo se aplican transformaciones y el analizador no “entiende” el código con lo cual es imposible que sepa cuando hay que importar dependencias.

Otro problema es que, como podemos ver en la figura 1, las funciones auxiliares no son traducidas, un programador Java debe traducir el código de Perl manualmente, lo cual impide crear códigos modulares.

El etiquetado en algunos casos es redundante, algunas etiquetas no aportan información que no pudiera extrapolarse directamente del código. Un ejemplo es la etiqueta <perl> que precede a cualquier etiqueta escrita en codigo Perl o <var> que indica que estamos tratando una variable. Cuanta más información saquemos del código fuente Perl, más rápido será el etiquetado.

Limitaciones de sintaxis, Perl tiene expresiones sintácticas que no están presentes en java, por ejemplo, aplicar estructuras de control a una sentencia o evaluar condiciones no booleanas. Las reglas de programación para Perldoop especifican las alternativas a los casos anteriores para que el programador pueda adaptar su código.

## Objetivos

* Traducción de un subconjunto de la sintaxis y código Perl a Java con solo introducir una serie de etiquetas que permiten definir tipos de variables, tamaño de arrays, porciones de código, etc.
* Soporte para todos los tipos de datos básicos, así como para Array y Hash.
* Conversión para los operadores matemáticos, lógicos, binarios y de cadenas.
* Casting automático entre tipos de datos básicos de forma transparente al usuario, transformando número a cadenas o enteros a flotantes, entre otros.
* Soporte para las estructuras de control que afectan tanto a bloques como a líneas, entre ellas podemos destacar if, while, for, foreach,etc.
* Llamadas a funciones nativas de Perl que se determinen esenciales o que sean muy usadas en la mayoría de los scripts.
* Definición de funciones personalizadas, aunque con alguna limitación, permitan separar el código en funciones para mejorar su legibilidad.
* Transformación de segmentos de código secuencial definidos en funciones mapper o reducer compatibles con Hadoop.
* Emulación básica de la aritmética de punteros de Perl, añadiendo la posibilidad de referenciar colecciones en vez de obligar a copiarlas.
* Sistemas de errores donde se informe al usuario de los errores producidos y las líneas donde han sido cometidos.
* Comprobaciones sobre el código para avisar de comportamientos no permitidos en Java como son el leer una variable sin inicializar o tener secciones muertas en código.
* Opciones de depuración para buscar errores en el análisis y facilitar correcciones en el traductor.

# Gestión del Alcance

En este capítulo se describe el resultado esperado como consecuencia de la realización de este proyecto.

## Descripción del alcance

El producto resultante del desarrollo de este proyecto será un traductor fuente-a-fuente que permita traducir un subconjunto limitado de la sintaxis de Perl a java. El software recibirá como entrada un script escrito en Perl y proporcionará como salida un código java que pueda ser copilado y ejecutado directamente.

La herramienta se presenta como la versión mejorada de la versión inicial, la sintaxis Perl será aumentada en la medida de lo posible para reconocer los patrones más usados en los scripts de procesamiento de lenguaje natural.

Para terminar, al igual que la versión original, el producto final incluirá un manual de uso donde se especifiquen las características y la sintaxis de la herramienta.

## Entregables

* **Perldoop 2.0**: Sofware terminado y listo para ser usado.
* **Memoria de proyecto**: Aspectos relacionados con la gestión del proyecto.
* **Presentación**: Documento para la presentación del TFG en su defensa.
* **Manual de uso**: Documento para familiarizar a los usuarios con el uso y el funcionamiento de la herramienta.

## Criterios de Aceptación

En base a los objetivos especificados anteriormente, se logrará la aceptación de producto final cuando se cumplan los siguientes criterios.

La herramienta deberá poder traducir cualquier script Perl escrito con la sintaxis dentro del marco especificado en este Proyecto. Por otro lado, en caso de que la herramienta no pueda efectuar la traducción por un error sintáctico o semántico, el usuario será informado de la causa del problema y además de la posición aproximada del mismo dentro del fichero.

En proyecto deberá contar con una serie de pruebas que aseguren que se cumple lo especificado previamente, así como que el código fuente de la herramienta está libre de errores sintácticos típicos en unos lenguajes de programación no copilados.

Por último, se considerará el proyecto aceptado cuando ambos tutores den el visto bueno tanto a esta documentación como a la herramienta de traducción.

## Restricciones del Proyecto.

Debido a la diferencia del nivel entre los lenguajes de programación Perl y java, se requiere complementar las sintaxis de perl con etiquetas en los scripts a traducir.

## Exclusiones del Proyecto

La finalidad de esta herramienta es proporcionar un método rápido de traducción de scripts Perl a Java, en ningún momento se intenta conseguir un traductor con las capacidades del intérprete del Perl. Para empezar la sintaxis orientada a objetos de Perl no será soportada por no ser necesaria.

## Supuestos del Proyecto

En este proyecto asumiremos que los scripts han sido probados y validados por el intérprete de Perl, pues no tiene sentido perder tiempo en solucionar problemas para los cuales ya existe una solución oficial por parte de los creadores del lenguaje.

# Casos de Uso

Los casos de uso definen los posibles escenarios de iteración de un usuario con el software. Los diagramas muestran el comportamiento del programa sin entrar en detalles de implementación.

En este caso el programa tiene una interfaz por consola, el programa toma todos sus argumentos en la invocación y no requiere de más intervención por el usuario. Por este motivo las acciones del usuario se ven reducidas a los argumentos de invocación.

En la figura 2 se muestra de manera gráfica las interacciones posibles entre el usuario y la herramienta traductora.

C:\Users\César\Desktop\casos de uso.png

Figura 2: Diagrama casos de uso

|  |  |
| --- | --- |
| Escala de importancia | |
| Alta | El caso de uso es crucial para el funcionamiento normal de la herramienta. |
| Media | El caso de uso puede resultar útil en la mayoría del uso de la herramienta, pero puede funcionar correctamente sin ellas. |
| Baja | El caso de uso complementa alguna acción sin mucha importancia pero que aumenten la versatilidad de la herramienta. |

|  |  |
| --- | --- |
| Caso de Uso 1 | |
| Nombre | Ejecutar Programa |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Alta |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece los parámetros para la traducción que será ejecutada mediante línea de comandos. | | 2 | El sistema analiza los parámetros y configura el traductor. | | 3 | El sistema abre el código fuente Perl a traducir. | | 4 | El sistema guarda el código Java generado. | |
| Postcondición | En código java ha sido generado y almacenado en el disco duro. |
| Excepciones | |  |  | | --- | --- | | Paso | Excepción | | 1 | 1-A: Si alguno de los parámetros es incorrecto, se aborta la ejecución y se muestra un mensaje de error. | | 2 | 2-A: Si no se puede abrir el script Perl a traducir, se aborta la ejecución y se muestra un mensaje de error. | | 3 | 3-A: Si no se puede guardar el código Java, se aborta la ejecución y se muestra un mensaje de error. | |

|  |  |
| --- | --- |
| Caso de Uso 2 | |
| Nombre | Abortar en error |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Baja |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema encuentra un error en el código fuente. | | 3 | El sistema aborta la ejecución y no sigue analizando. | |
| Postcondición | Ninguna |
| Excepciones | Ninguna |

|  |  |
| --- | --- |
| Caso de Uso 3 | |
| Nombre | Interpretar comentario |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Media |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema encuentra un comentario en el código fuente | | 3 | El sistema comprueba si el comentario es valido | | 4 | El sistema traduce el comentario a java | |
| Postcondición | En código java contiene los comentarios escritos en Perl. |
| Excepciones | |  |  | | --- | --- | | Paso | Excepción | | 1 | 1-A: Si el comentario está escrito en medio de una sentencia, aborta la ejecución y se muestra un mensaje de error. | |

|  |  |
| --- | --- |
| Caso de Uso 4 | |
| Nombre | Emular paréntesis |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Media |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema encuentra una función sin paréntesis en sus argumentos. | | 3 | El sistema comprueba el número de parámetros de la función. | | 4 | El sistema añade manualmente los paréntesis contando las comas. | |
| Postcondición | Ninguna |
| Excepciones | |  |  | | --- | --- | | Paso | Excepción | | 1 | 1-A: Si la función no existe, no se añaden los paréntesis y se delega el error al analizador. | | 2 | 2-A: Si la función tiene menos parámetros de los requeridos, solo se añaden los que hay y se delega el error al analizador. | |

|  |  |
| --- | --- |
| Caso de Uso 5 | |
| Nombre | Modo depuración |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Baja |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema empieza a realizar la traducción | | 3 | El sistema imprime las tareas que va realizando | |
| Postcondición | Ninguna |
| Excepciones | Ninguna |

|  |  |
| --- | --- |
| Caso de Uso 6 | |
| Nombre | Optimizar código |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Alta |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema invierte tiempo en aplicar patrones al código para generar un código más legible y optimizado. | |
| Postcondición | Ninguna |
| Excepciones | Ninguna |

|  |  |
| --- | --- |
| Caso de Uso 7 | |
| Nombre | Buscar código muerto |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Media |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema comprueba si existen sentencias inalcanzables. | |
| Postcondición | Ninguna |
| Excepciones | |  |  | | --- | --- | | Paso | Excepción | | 1 | 1-A: Si encuentra sentencias inalcanzables, se informa al usuario con un error. | |

|  |  |
| --- | --- |
| Caso de Uso 8 | |
| Nombre | Generar main |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Baja |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema crea la función de arranque main con el código global del script. | |
| Postcondición | El código java contiene la función main. |
| Excepciones | Ninguna |

|  |  |
| --- | --- |
| Caso de Uso 9 | |
| Nombre | Cambiar ruta de salida |
| Descripción | El sistema deberá comportarse como se describe en el siguiente caso de uso cuando un usuario trate de ejecutar el programa. |
| Importancia | Media |
| Precondición | Ninguna |
| Secuencia Normal | |  |  | | --- | --- | | Paso | Acción | | 1 | El usuario establece el parámetro | | 2 | El sistema genera la traducción | | 3 | El sistema guarda los ficheros en el directorio especificado | |
| Postcondición | Los archivos generados se encuentran en el directorio especificado. |
| Excepciones | |  |  | | --- | --- | | Paso | Excepción | | 1 | 1-A: Si el directorio no existe o no se tiene acceso, aborta la ejecución y se muestra un mensaje de error. | |

# Análisis de requisitos

En las reuniones con los tutores del proyecto se fueron definiendo una serie de requisitos que el software final deberá de cumplir. En esta sección se clasificarán los requisitos según el estándar IEEE830[7].

Los requisitos podemos clasificarlos atendiendo a dos criterios.

* Funcionales y no funcionales.
* Según su importancia:
  + Vital: Son necesario para la aceptación del proyecto.
  + Importante: El sistema puede funcionar correctamente sin ellos, pero el usuario esperara disponer de ellos.
  + Quedaría bien: Añaden calidad al producto final.

## Requisitos funcionales

Los requisitos funcionales definen las funciones del sistema y sus componentes, a continuación, se enumeran los encontrados en este proyecto.

|  |  |
| --- | --- |
| RF-1 | |
| Nombre | Implementación de interfaz por consola |
| Descripción | La interacción con el usuario se llevará a través de una interfaz por consola, el usuario definirá los parámetros y luego ejecutará el programa. |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-2 | |
| Nombre | Definir scripts a traducir |
| Descripción | El usuario puede definir un conjunto de scripts que desea traducir a java. |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-3 | |
| Nombre | Definir carpeta de salida |
| Descripción | El usuario puede elegir una carpeta de salida distinta de la actual. |
| Importancia | Importante |

|  |  |
| --- | --- |
| RF-4 | |
| Nombre | Traducir variables básicas, array, hash y descriptores de fichero. |
| Descripción | El sistema debe poder interpretar las variables en el código, así mismo también inicializarlas y acceder a sus dimensiones. |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-5 | |
| Nombre | Traducción de operaciones |
| Descripción | El sistema deberá traducir las operaciones de tipo: Matemáticas, Lógicas, Binarias, Concatenación, Repetición y Asignación. |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-6 | |
| Nombre | Casting automático |
| Descripción | El sistema deberá convertir entre los distintos tipos de datos de forma transparente para el usuario. |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-7 | |
| Nombre | Traducción de estructuras de control |
| Descripción | El Sistema deberá traducir las siguientes estructuras de control:  if-elsif-else, while, do while, for, foreach |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-8 | |
| Nombre | Traducción funciones nativas Perl |
| Descripción | El sistema deberá traducir las siguientes funciones nativas Perl:  chomp, chop, close, defined, delete, die, each, exists, exit, join  keys, lc, lcfirst, length, open, pop, print, push, say, shirft  sort, splice, split, substr, system, uc, ucfirst, unshift, values |
| Importancia | Quedaría Bien |

|  |  |
| --- | --- |
| RF-9 | |
| Nombre | Declaración de funciones propias en Perl |
| Descripción | El usuario podrá definir funciones propias dentro del código Perl. |
| Importancia | Importante |

|  |  |
| --- | --- |
| RF-10 | |
| Nombre | Generar código Hadoop |
| Descripción | El sistema podrá convertir código secuencial definidos en funciones mapper o reducer compatibles con Hadoop. |
| Importancia | Vital |

|  |  |
| --- | --- |
| RF-11 | |
| Nombre | Implementar sistema de error |
| Descripción | El sistema deberá avisar al usuario cuando encuentre un error en el código fuente. |
| Importancia | Importante |

|  |  |
| --- | --- |
| RF-12 | |
| Nombre | Implementar sistema de depuración |
| Descripción | El sistema deberá poder informar de su funcionamiento interno para facilitar el mantenimiento de la misma. |
| Importancia | Quedaría bien |

|  |  |
| --- | --- |
| RF-13 | |
| Nombre | Verificar código generado |
| Descripción | El sistema deberá buscar irregularidades en el código java para luego poder avisar al usuario. |
| Importancia | Usuario |

|  |  |
| --- | --- |
| RF-14 | |
| Nombre | Optimizar código generado |
| Descripción | El sistema deberá intentar optimizar el código final aun a coste del tiempo de traducción |
| Importancia | Quedaría bien |

|  |  |
| --- | --- |
| RF-15 | |
| Nombre | Traducir comentarios Perl |
| Descripción | El sistema deberá conservar los comentarios del programador en el código fuente copiándolos en el código final. |
| Importancia | Quedaría bien |

|  |  |
| --- | --- |
| RF-16 | |
| Nombre | Formatea código java |
| Descripción | El sistema deberá producir un código legible para el usuario, es importante que el código final este correctamente identado. |
| Importancia | Quedaría bien |

|  |  |
| --- | --- |
| RF-17 | |
| Nombre | Generar paréntesis funciones |
| Descripción | El sistema debería permitir que las funciones especificadas sin paréntesis fueran traducidas sin la necesidad de añadírselos. |
| Importancia | Quedaría bien |

|  |  |
| --- | --- |
| RF-17 | |
| Nombre | Generar main automático |
| Descripción | El sistema debería generar un punto de ejecución para el programa en caso de querer generar código secuencial. |
| Importancia | Quedaría bien |

## Requisitos no funcionales

Los requisitos no funcionales definen características del sistema que pueden usarse para validar la aplicación una vez terminada. Por lo tantos se refirieren a todos los requisitos que no especifican la información a guardar ni la función a realizar, solo las características.

|  |  |
| --- | --- |
| RNF-1 | |
| Nombre | Portabilidad |
| Descripción | La aplicación deberá funcionar y comportarse del mismo modo en Windows, Linux y Mac. |
| Importancia | Importante |

|  |  |
| --- | --- |
| RNF-2 | |
| Nombre | Paradigma orientado a objetos |
| Descripción | La implementación del código fuente deberá desarrollarse con el paradigma orientado a objetos |
| Importancia | Importante |

|  |  |
| --- | --- |
| RNF-3 | |
| Nombre | Extensibilidad |
| Descripción | La aplicación deberá estar diseñada de forma que permita añadir nuevas funcionales sin afectar a las existentes. |
| Importancia | Importante |

|  |  |
| --- | --- |
| RNF-4 | |
| Nombre | Tecnología apropiada |
| Descripción | La aplicación deberá estar implementada usando la tecnología apropiada para traductores, es decir analizador léxico, sintáctico y semántico, en vez de aplicar expresiones regulares con trasformaciones directas. |
| Importancia | Vital |

# Gestión del Proyecto

En esta fase de gestión del proyecto se abordan las medidas necesarias para realizar un control sobre el proyecto durante su ciclo de vida. El principal objetivo de este apartado es cumplir con el alcance dentro de las restricciones temporales, económicas y de calidad.

## Metodología

Para evitar que el proyecto fracase, antes de transformar los requisitos es necesario la elección de una metodología que se adecue al proyecto y además permita satisfacer la planificación del mismo. La importancia de la metodología recae en dar solución a la hora de elegir la forma de estructura, planificar y gestionar el proceso de desarrollo de una aplicación concreta. Para su elección o debemos analizar qué modelo de clico de vida se adapta mejor en nuestro caso.

El proyecto tiene un carácter individual, de modo que no existe la necesidad de coordinarse con un equipo o necesidad de reuniones entre los integrantes, lo importante es profundizar lo suficiente en el diseño o en la ingeniería de software lo necesario para aclarar lo que se va a desarrollar y el cómo sin profundizar en elementos que no aporten soluciones aplicables.

Por otro lado, el proyecto presenta una fase importante de investigación, y a pesar de que los requisitos no cambiaran durante el desarrollo del proyecto, es posible que sean necesarios cambios según se valla profundizando en la muy permisiva sintaxis de Perl. Por este motivo es necesario un ciclo de desarrollo ágil y cíclico que permita adaptar el código a los posibles cambios.

Dadas las características anteriores se necesita una metodología que permita: flexibilidad en el desarrollo, procesos simples de terrollo y un proceso cíclico de revisión para poder aplicar correcciones en caso de ser necesario y poder cumplir con la fecha de entrega.

Por los motivos expuestos se ha seleccionado como metodología de desarrollo la Programación Extrema.

## Programación extrema

La Programación extrema [9] es una metodología ágil de desarrollo de software cuyo principal objetivo es crear un software de calidad centrándose en el desarrollo de código. La filosofía se centra en un ciclo de desarrollo y prueba continuos, pudiendo así adaptarse a cambios de forma dinámica cuando cambian los requisitos o el ritmo de desarrollo es insuficiente. En este caso, como se dijo anteriormente, los requintos se mantendrán fijos durante todo el ciclo de visa, pero la sintaxis de Perl puede ocasionar que sean necesarias modificaciones para que la traducción sea válida. Por otro lado, la fecha de entrega del proyecto es inamovible y dada la complejidad y cantidad de código necesaria para implementar un traductor es importante invertir la mayor parte del tiempo posible en el desarrollo en vez de creando documentación. Este ciclo de vida permite utilizar el código como propia documentación cuando sea posible, reduciendo así la carga de los procesos de ingeniería de software.

El modelo de ciclo de vida software está pensado para equipos de trabajos pequeños o medianos y este caso no podría ser más pequeño, además del seguimiento de buenas prácticas. Otras de sus características es que se presta especial atención a la realimentación, el trabajo se divide en ciclos cortos, la validación se lleva a cabo mediante pruebas unitarias, diseños simples que permitan cumplir los requitos de forma fácil, etc.

## Gestión de la configuración

Se entiende por configuración en el desarrollo del software, todas las características descritas en la documentación técnica o que son alcanzadas por el producto. La gestión de configuración, tiene como objetivo establecer y mantener la integridad de todo producto creado durante el desarrollo a través de una serie de tareas. En primer lugar, hay que definir los elementos sobre los que se aplica la gestión de la configuración, luego hay que definir un proceso de cambios y por ultimo auditorias de configuración para verificar el cumplimiento del mismo.

Un dato a tener en cuenta es que la única finalidad de la gestión de la configuración en este proyecto es mantener un control de versiones, ya que en ningún momento existirá otra persona distinta del alumno trabajando sobre el proyecto. Esto se ha tenido en cuenta la hora de realizar esta gestión de la configuración, adaptando solo las partes necesarias, aligerando el desarrollo de este proceso para evitar malgastar el tiempo.

Cabe añadir que la función del proceso de gestión de la configuración en este proyecto es meramente la de mantener un control de versiones, ya que no se realizará trabajo de forma concurrente por parte de varias personas: el alumno será el único que modifique los elementos de configuración. Esto se ha tenido en cuenta a la hora de realizar la planificación de la gestión de configuración, adaptándola a los fines que realmente tendrá, aligerando el peso de este proceso durante el desarrollo del proyecto.

## Identificación de elementos de configuración

Los elementos de configuración detectados son los siguientes:

* Código fuente
* Código de pruebas
* Memoria de proyecto
* Manual de la aplicación
* Presentación del proyecto

Todos los elementos, excepto el código fuente y pruebas, deberán seguir el siguiente patrón:

<Tipo de documento>\_<AAMMDD>\_[v<versión>].<extensión>

Donde el tipo de documento pude ser:

|  |  |  |  |
| --- | --- | --- | --- |
| Tipo de documento | Descripción | Código | Extensión |
| Memoria de proyecto | Documento base con información acerca del proyecto | MEN | .pdf |
| Manual de la aplicación | Manual de usuario | MAN | .pdf |
| Presentación del proyecto | Diapositivas para la exposición del proyecto | PDP | .ptt |

Tabla 1: Tipos de documento

La fecha estará en formato AAMMDD corresponde a la fecha de modificación del elemento y la versión hace referencia a la versión interna del mismo, con cada modificación el número de versión incrementa.

Por último, la extensión hace referencia al formato que en principio siempre está presente en sistemas tipo Windows.

## Herramienta gestión de la configuración

La elección de la herramienta de gestión de la configuración es un paso muy importante antes de iniciar un proyecto. La herramienta debe conservar las distintas versiones de los elementos y además permitir el trabajo de múltiples personas sobre el mismo trabajo.

Para alojar y controlar las distintas versiones del código fuente y los elementos, se ha considerado la elección de un repositorio tipo GIT. El repositorio deberá estar alojado en un servidor en la nube para asegurar la persistencia del trabajo en caso de que el tipo de trabajo se estropee y además permitir a los tutores del proyecto poder seguir el desarrollo del mismo.

Por todas estas razones se ha optado por alojarlo en la web BitBucket[9], esta web ofrece repositorios privados donde pueden trabajar 3 personas de forma gratuita, lo cual en este caso es el numero perfecto.

## Planificación temporal

En este apartado se aborda la planificación temporal comentado por un EDT (Estructura de descomposición de Trabajo) divididos en varios apartados según su especificación. El EDT es una representación jerárquica del proyecto en entregables o áreas de trabajo para determinar el ámbito del trabajo y poder planificar los objetivos propuestos.

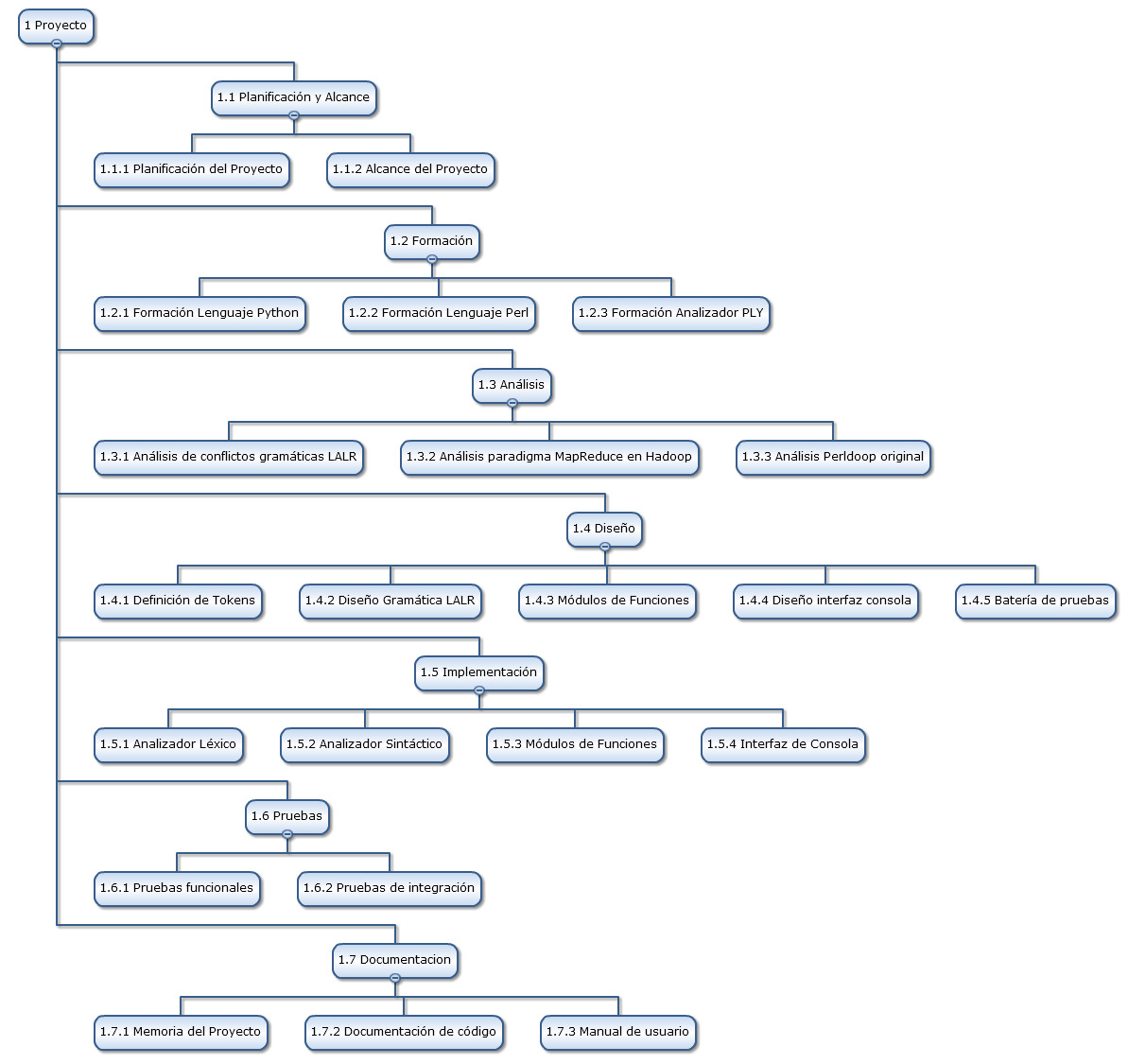
**

Figura 3: Estructura de descomposición de trabajo

## Planificación y Alcance

En esta etapa se realiza un estudio sobre el proyecto a realizar, definiendo los objetivos, requisitos, casos de uso e identificando los posibles riegos que pueden afectar al proyecto. Esta fase es el punto de partida en la que se basan las fases posteriores.

* Tiempo estimado: 1 Semana
* Rol: Analista-Programador

## Formación

Para la realización del TFG es necesario el conocimiento en los lenguajes de Programación Java, Python y Perl. El primero se estudia con bastante profundidad en la carrera, pero los dos últimos requieren de una mejor formación.

Además, es necesario familiarizarse con la tecnología de construcción de traductores en Python llamada PLY [10].

* Tiempo estimado: Media Semana
* Rol: Analista-Programador

## Análisis

Para poder diseñar correctamente un traductor es necesario conocer en profundidad las posibilidades de la gramática antes de comenzar con su diseño. Por otro lado, es necesario familiarizarse con el paradigma de Hadoop y analizar detenidamente la versión anterior de la herramienta para no repetir los mismos errores.

* Tiempo estimado: Media Semana
* Rol: Analista-Programador

## Diseño

El proyecto requiere el diseño de 5 módulos principales:

1. **Tokens:** La primera parte del diseño consiste en definir los tokens léxicos de Perl que formaran parte de nuestro proyecto.
2. **Gramática**: La gramática es la parte crucial de este proyecto, se debe definir una gramática lo suficientemente amplia como para cumplir los objetivos del proyecto.
3. **Módulos de funciones**: Las funciones de generación de código deben diseñarse de forma que se pueda reutilizar el mayor código posible.
4. **Interfaz de consola**: La interacción con el usuario en la parte más importante de la herramienta, una mala interfaz puede perjudicar la opinión sobre la herramienta.
5. **Pruebas**: Es necesario validar la herramienta con los mínimos casos de pruebas que permitan eliminar los máximos errores posibles.

* Tiempo estimado: 1 Semana y media
* Rol: Analista-Programador

## Implementación

Esta es la fase más larga en toda la planificación del proyecto. Consiste en la implementación de todo el proyecto empezando por el analizador léxico hasta la generación de código.

* Tiempo estimado: 8 Semanas
* Rol: Analista-Programador

## Pruebas

Esta parte engloba el desarrollo de las pruebas especificadas en el apartado de diseño, además de las correcciones sobre el código fuente en caso de encontrarse algun error.

* Tiempo estimado: 2 Semanas
* Rol: Analista-Programador

## Documentación

La última parte de la planificación que se ira realizando de manera incremental según se avanza en las demás fases, engloba la documentación del proyecto, el manual de usuario y la creación de este documento de memoria de proyecto.

* Tiempo estimado: 2 Semanas
* Rol: Analista-Programador

## Cronograma

La tabla siguiente muestra la planificación temporal del proyecto por semanas de igual forma que se hizo en el anteproyecto.

|  |  |
| --- | --- |
| Estimación temporal (en semanas) | |
| Planificación y Alcance | **0,98** |
| Planificación del Proyecto | 0,61 |
| Alcance del Proyecto | 0,37 |
| Formación | **0,56** |
| Formación Lenguaje Python | 0,22 |
| Formación Lenguaje Perl | 0,22 |
| Formación Analizador PLY | 0,11 |
| Análisis | **0,56** |
| Análisis de conflictos gramáticas LALR | 0,11 |
| Análisis paradigma MapReduce en Hadoop | 0,22 |
| Análisis Perldoop original | 0,22 |
| Diseño | **1,50** |
| Definición de Tokens | 0,34 |
| Diseño Gramática LALR | 0,56 |
| Módulos de Funciones | 0,11 |
| Diseño interfaz consola | 0,11 |
| Batería de pruebas | 0,37 |
| Implementación | **7,48** |
| Analizador Léxico | 0,75 |
| Analizador Sintáctico | 1,12 |
| Módulos de Funciones | 5,98 |
| Interfaz de Consola | 0,37 |
| Pruebas | **2,06** |
| Pruebas funcionales | 1,31 |
| Pruebas de integración | 0,75 |
| Documentación | **1,87** |
| Memoria del Proyecto | 1,12 |
| Documentación de código | 0,37 |
| Manual de usuario | 0,37 |
| Total | **15** |

Tabla 2: Cronograma

## Diagrama de Gantt

El diagrama de Gantt ofrece una representación gráfica de la planificación de las actividades identificados en la Estructura de Descomposición de Trabajo. La granularidad de esta descomposición dependerá de la estimación del nivel necesario para una estimación fiable.

Como recurso asignado al proyecto, solo tenemos al alumno que trabajara todos los días durante 4 horas para terminar el proyecto en 15 semanas. El proyecto se desarrollará desde la última semana de marzo hasta principio de Julio.

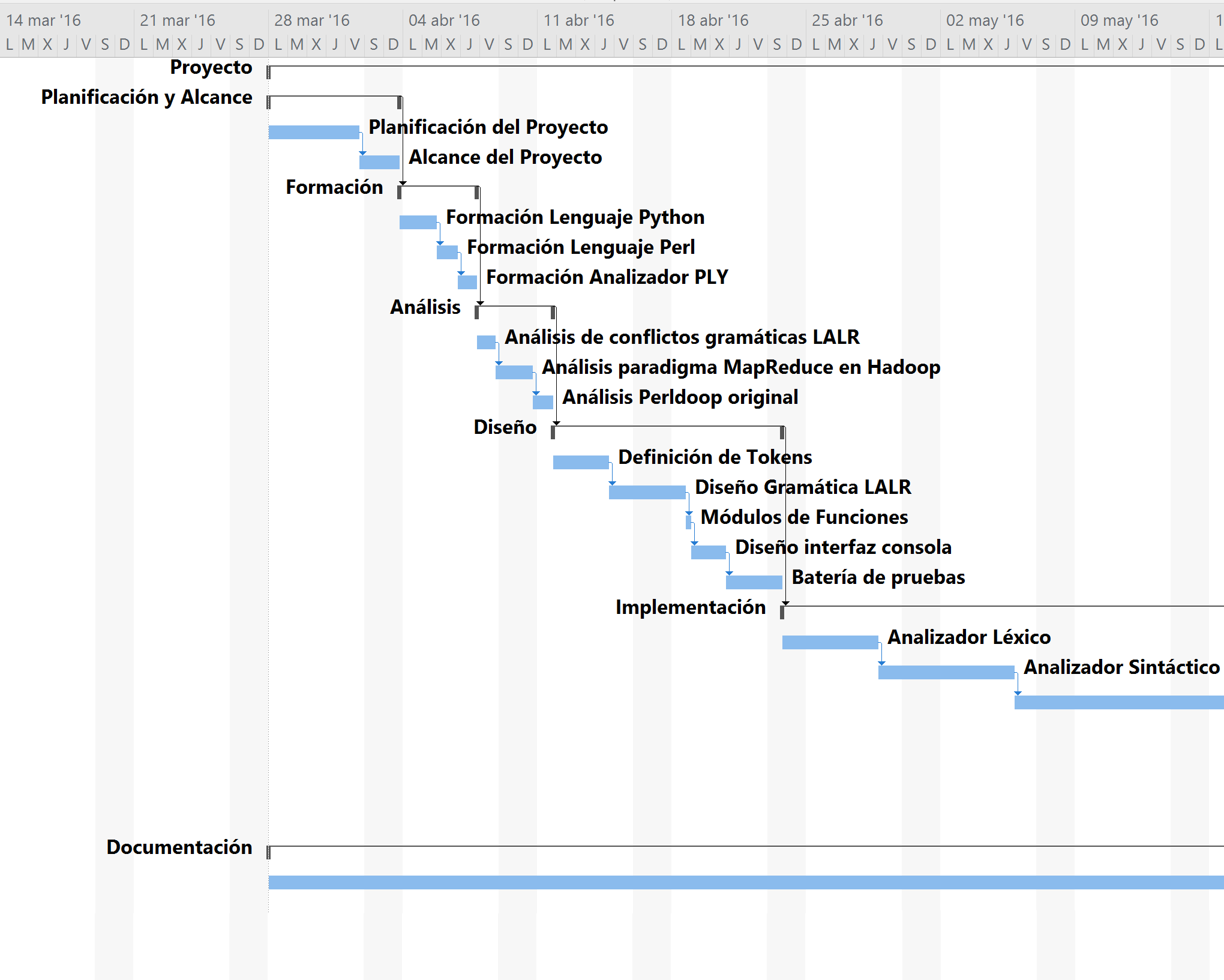


Figura 4: Diagrama Gantt parte 1

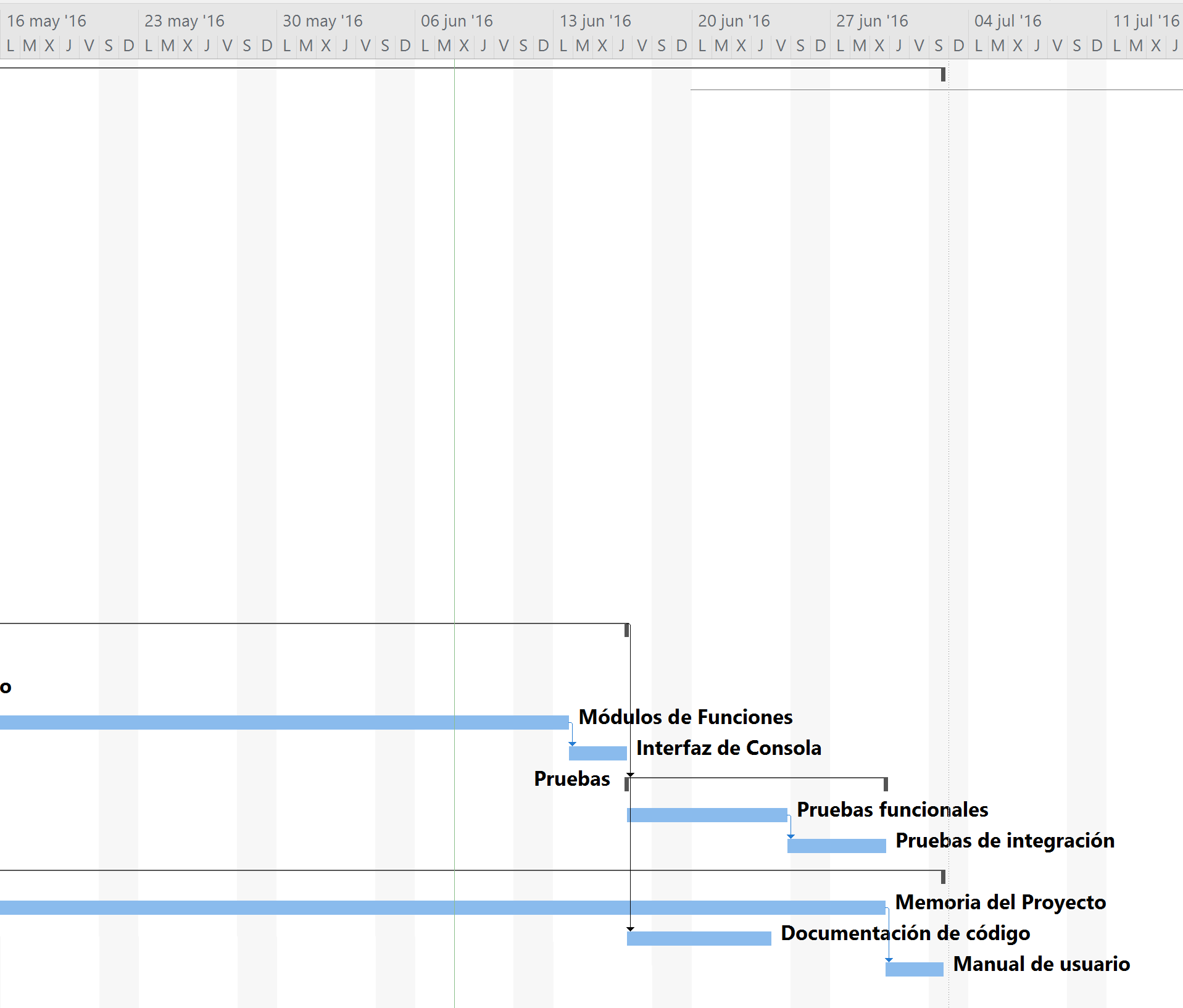


Figura 5: Diagrama Gantt parte 2

## Gestión de riesgos

Los riegos son sucesos inciertos de naturaleza positiva o negativa que pueden manifestarse durante el desarrollo del proyecto, en este caso solo identificaremos los riesgos de carácter negativo. Esta tiene como objetivo identificar, analizar y cuantificar los posibles riesgos que puedan presentarse. Una vez identificados se procederá a clasificarlos siguiendo la siguiente clasificación:

Repercusión que tendría un riego en caso de darse sobre las posibilidades de completar el proyecto de forma correcta en el tiempo planificado.

|  |  |
| --- | --- |
| ESCALA DE IMPACTO | |
| Alta | El impacto podría afectar gravemente a la calidad final del producto o causar un retraso importante en la planificación. |
| Media | El impacto podría afectar levemente la calidad final del producto o causar un retraso leve en la planificación. |
| Baja | El impacto podría afectar a la planificación pero no lo suficiente como para retrasar la fecha de finalización. |

Estimación de la probabilidad que tiene un riego en aparecer.

|  |  |
| --- | --- |
| ESCALA DE PROBABILIDAD | |
| Alta | Probabilidad igual o mayor que 70% |
| Media | Probabilidad entre el 30% y 70% |
| Baja | Probabilidad igual o menor que 30% |

## Especificación

Se procede a la identificación y clasificación de riegos siguiendo la metodología descrita en el PMBOK [6].

**Riesgos del Proyecto**

Se definen los riesgos del proyecto a aquellos riegos que afectan a la gestión del proyecto, en caso de producirse, afectaran a la planificación temporal del proyecto y al coste del mismo.

|  |  |
| --- | --- |
| ID | RP-1 |
| Nombre | Retraso respecto a la planificación |
| Descripción | Puede darse el caso de que alguna tarea se haya planificado mal y esto casusa un retraso en la fecha de entrega. |
| Probabilidad | Media |
| Impacto | Alto |
| Plan de prevención | Realizar reuniones periódicas con los tutores de proyecto para vigilar el avance del proyecto. |
| Plan de continencia | Aumentar el número de horas diarias hasta alcanzar la planificación inicial. |

|  |  |
| --- | --- |
| ID | RP-2 |
| Nombre | Malgastar tiempo documentando |
| Descripción | El caso de cambios en el desarrollo del proyecto, puede ser necesario adaptar la documentación perdiendo tiempo de desarrollo. |
| Probabilidad | Media |
| Impacto | Medio |
| Plan de prevención | No documentar en profundidad las partes sujetas a cambios hasta asegurar su permanencia. |
| Plan de continencia | Aumentar el número de horas diarias para reparar el documento. |

|  |  |
| --- | --- |
| ID | RP-3 |
| Nombre | Enfermedad |
| Descripción | Si el alumno cae enfermo habrá un periodo de tiempo donde no se pueda seguir la planificación. |
| Probabilidad | Baja |
| Impacto | Medio |
| Plan de continencia | Restructurar la planificación para repartir las horas perdidas en el resto de días. |

**Riesgos Técnicos**

Los riegos técnicos, son aquellos que pueden amenazar la calidad final de producto y el tiempo de desarrollo del software que pueden repercutir en la planificación temporal.

|  |  |
| --- | --- |
| ID | RT-1 |
| Nombre | Diseño incorrecto de la gramática del analizador sintáctico |
| Descripción | La gramática creada en la fase de diseño contiene conflictos y no puede ser implementada. |
| Probabilidad | Media |
| Impacto | Alto |
| Plan de prevención | Diseñar la gramática directamente sobre el analizador para evitar conflictos. |
| Plan de continencia | Rediseñar la gramática hasta eliminar los conflictos. |

|  |  |
| --- | --- |
| ID | RT-2 |
| Nombre | Tiempo de depuración prolongado |
| Descripción | La depuración del traductor consume mucho más tiempo del esperado con la aparición de muchos errores. |
| Probabilidad | Alta |
| Impacto | Alto |
| Plan de prevención | Implementar las pruebas antes de la herramienta para evitar la mayoría de los fallos. |
| Plan de continencia | Invertir horas extras en pruebas hasta volver a la planificación. |

|  |  |
| --- | --- |
| ID | RT-3 |
| Nombre | Perdida del trabajo de un día |
| Descripción | Un fallo en el sistema evita el acceso a los datos almacenados en el ordenador. |
| Probabilidad | Baja |
| Impacto | Alto |
| Plan de prevención | Comenzar a trabajar con el repositorio y actualizarlo al menos una vez al día. |
| Plan de continencia | Repetir el trabajo lo antes posible para recordar exactamente lo que se ha perdido. |

|  |  |
| --- | --- |
| ID | RT-4 |
| Nombre | Requisito imposible de implementar |
| Descripción | Un requisito especificado para el proyecto es imposible de Implementar. |
| Probabilidad | Bajo |
| Impacto | Alto |
| Plan de prevención | Diseño preliminar de una aproximación en la implantación para cada requisito del sistema. |
| Plan de continencia | Reunión con los tutores del proyecto para buscar solución al problema. |

|  |  |
| --- | --- |
| ID | RT-5 |
| Nombre | Traducciones incorrectas |
| Descripción | Las traducciones de la herramienta no son baldías tanto por no poder ser copiladas como por no hacer lo mismo que el código fuente. |
| Probabilidad | Medio |
| Impacto | Alto |
| Plan de prevención | Probar cada traducción segunda se va implementando de forma unitario y en conjunto con el resto. |
| Plan de continencia | Esperar a la fase de pruebas para corregir los errores de traducción. |

|  |  |
| --- | --- |
| ID | RT-6 |
| Nombre | Errores sintácticos en Python |
| Descripción | Los lenguajes interpretados pueden contener errores sintácticos en líneas que no son ejecutadas. |
| Probabilidad | Alta |
| Impacto | Bajo |
| Plan de prevención | Verificar que todas las líneas son interpretadas. |
| Plan de continencia | Corregir los errores en cuanto aparezcan mientras se continua la implementación o con el plan e pruebas. |
| ID | RT-7 |
| Nombre | Interfaz de consola poco usable |
| Descripción | La interfaz de consola es poco intuitiva y es difícil familiarizarse con ella |
| Probabilidad | Baja |
| Impacto | Alto |
| Plan de prevención | Pedir opinión a otras personas acerca de la interfaz y realizar cambios con sus sugerencias. |
| Plan de continencia | Añadir un manual para mitigar una interfaz complicada. |

## Análisis de costes

Una vez terminada la planificación temporal del proyecto, es necesario estimar los costes que acarrearía el proyecto en una situación real. En esta fase se incluye el coste humano (el alumno) así como las herramientas y recursos necesario para la creación del proyecto. Todos los programas utilizados durante el desarrollo están disponibles de manera gratuita en internet y pueden descargarse desde su página oficial. El único material usado es un ordenador portátil propiedad del alumno con lo cual el coste total de proyecto viene dado por las horas invertidas por el alumno en la realización del mismo.

Para realizar el cálculo de la mano de obra se han consultado varias webs como InfoJobs [11] o la calculadora de contratos de la USC para buscar el sueldo medio de un ingeniero informático en proyectos de este tipo. Se ha estipulado un sueldo aproximado de 17,2€ la hora.

En la siguiente tabla se muestra una división de costes por cada tarea del proyecto.

|  |  |
| --- | --- |
| **Tarea** | Coste |
| **Planificación y Alcance** | **450,90 €** |
| Planificación del Proyecto | 280,66 € |
| Alcance del Proyecto | 170,24 € |
| **Formación** | **253,06 €** |
| Formación Lenguaje Python | 101,22 € |
| Formación Lenguaje Perl | 101,22 € |
| Formación Analizador PLY | 50,61 € |
| **Análisis** | **253,06 €** |
| Análisis de conflictos gramáticas LALR | 50,61 € |
| Análisis paradigma MapReduce en Hadoop | 101,22 € |
| Análisis Perldoop original | 101,22 € |
| **Diseño** | **685,55 €** |
| Definición de Tokens | 156,43 € |
| Diseño Gramática LALR | 257,66 € |
| Módulos de Funciones | 50,61 € |
| Diseño interfaz consola | 50,61 € |
| Batería de pruebas | 170,24 € |
| **Implementación** | **3.782,02 €** |
| Analizador Léxico | 345,08 € |
| Analizador Sintáctico | 515,31 € |
| Módulos de Funciones | 2.751,40 € |
| Interfaz de Consola | 170,24 € |
| **Pruebas** | **947,81 €** |
| Pruebas funcionales | 602,73 € |
| Pruebas de integración | 345,08 € |
| **Documentacion** | **855,79 €** |
| Memoria del Proyecto | 515,31 € |
| Documentación de código | 170,24 € |
| Manual de usuario | 170,24 € |
| **Total** | **7.228,17 €** |

En resumen, si tenemos en cuenta las otras áreas:

|  |  |
| --- | --- |
| Fuente | Coste |
| Estación de trabajo | 0.0€ |
| Software utilizado | 0.0€ |
| Repositorio web | 0.0€ |
| Mano de obra | 7.228,17 € |
| Total | **7.228,17 €** |

# Herramientas y Tecnologías

En este apartado se describirán las diferentes herramientas y tecnologías utilizadas para la creación de este proyecto, entendemos por herramienta a cualquier aplicación usada para crear algún elemento del proyecto y las tecnologías todo aquello en que se apoya el proyecto para funcionar.

## IDEs de desarrollo

* **Eclipse** [12]: Se ha utilizado este IDE para el desarrollo del código fuente del traductor, eclipse tiene un plugin muy completo para el desarrollo de python3, al contrario que netbeans. Todo el código fuente y la depuración del mismo fueron realizados con el.
* **Netbeans** [13]: Se ha utilizado este IDE para el desarrollo de una librería auxiliar java que simplifica el código de traducción. La razón de no haber usado eclipse es la costumbre de haber realizado muchos proyectos java en él y conocer muchos atajos rápidos para un desarrollo lo mas rápido posible. Si no fuera porque no existe un plugin python3 para este IDE, todo el desarrollo se hubiera realizado con él.

## Tecnologías

* **Python**: Se ha optado la realización de traductor en este lenguaje de programación por ser un lenguaje de alto nivel, rápido de programar y que ofrece portabilidad entre distintas plataformas. El proceso de traducción requiere concatenar y dividir constantemente arrays y cadenas, lo cual se puede hacer con un simple operador. Otros lenguajes como Java o C no tienen soporte tan directo y se necesitaría más tiempo de desarrollo.
* **PLY** [10]: Entre los distintos analizadores léxicos y sintácticos disponibles para Python, se ha optado por PLY debido a su simplicidad y a su integración directa en el código Python sin necesidad de invocar a un preprocesador que transforme su código a Python. El analizador sintáctico de PLY es ascendente y permite definir gramáticas LALR(1) lo cual es suficiente para cumplir los objetivos del proyecto.
* **Perl**: En el desarrollo de un traductor Perl-java es imprescindible contar con el intérprete Perl para probar ficheros de ejemplo y verificar que tanto ellos como sus tradiciones ofrecen las mismas salidas una vez ejecutados.

## Aplicaciones

* **Notepad++**: Aunque para el desarrollo de código se han usado IDE, los scripts Perl y el fichero que contiene la definición de la gramática se han creado con este editor de texto. Por otro lado, el editor contiene funciones muy útiles como la de grabar y repetir acciones que ayuda a replicar y maquetar código de formas que el IDE no soporta.
* **SourceTree** [14]: Como ya se ha especificado en el apartado de gestión de la configuración, para el control de versiones del código fuente se ha optado por un repositorio GIT, alojado en la página BitBucket. Para el acceso y las gestiones se han considerado diversos clientes gráficos, pero al final se ha optado por SourceTree por tener una interfaz muy completa y contar con el apoyo de muchos usuarios. La herramienta permite realizar todas las tareas necesarias para el proyecto, permite realizar commit de forma rápida, ver el historial de commits en forma de árbol (o línea en nuestro caso) , comparar diferentes versiones del mismo fichero y volver atrás seleccionando el nodo en el árbol de commits.
* **Microsoft Office**: Se ha utilizado la suite de ofimática para la creación de la memoria del proyecto, más concretamente las versiones Word, Excel y Project. Se ha optado elegir esta versión frente a las versiones libres por ser mucho más completas en cuanto a opciones y además ya estaban instaladas en el sistema debido a su uso en las prácticas de la carrera. Aunque se trata de un programa de pago, esto no implica ningún coste debido a que se amortizaron durante toda la carrera y la versión Project se usó la versión de demostración suficiente para este proyecto.
* **Draw.io** [15]: Para crear todos los diagramas de la memoria del proyecto se ha optado por esta herramienta online que permite crear todo tipo de diagramas y además almacenarlos en la nube.
* **WBS TooL**: Esta herramienta web se ha usado para la creación de la Estructura de descomposición de trabajo, requerido como porte importante de la gestión del proyecto. Se ha optado por usar esta herramienta por su familiarización en las asignaturas de Ingeniería del Software y gestión de proyecto. La herramienta ofrece todas las funciones necesarias para generar un EDt de forma gráficas y luego permite exportar como xml para poder usarlo como base en la construcción de un diagrama de Gantt.
* **Microsoft Project**: La gestión del proyecto requiere de la construcción de un diagrama Gantt con la planificación estipulada por el proyecto que abarque las horas de un trabajo de fin de grado. Se ha optado por la versión de Microsoft frente a Project Libre debido a que al contrario que en las prácticas de las asignaturas, no podemos tolerar fallos que suelen ocurrir con la versión libre tales como error en los cálculos después de guardar y volver a cargar un proyecto. Además, el proyecto tiene un horario distinto del laboral donde se requiere definir un calendario nuevo, esta característica solo está disponible en la versión Microsoft porque la versión libre contiene un bug que ignora los calendarios.

# Arquitectura del software

En este apartado se discute acerca de la arquitectura en la que se ha basado la construcción del software, identificando cada uno de los componentes que lo forman, así como una breve descripción del papel de cada uno.

Los componentes son: Una interfaz de consola, el analizador léxico, el analizador sintáctico, el comprobar y generador de código, el sistema de mensajes, el sistema de errores y el propio sistema de ficharos del sistema operativo. La disposición entre ellos puede verse en el diagrama de la Figura 6.

D:\Users\Informatica\cuarto\Proyecto\arquitectura.png

Figura 6: Arquitectura del sistema

## Interfaz línea comandos

Para la interacción con el usuario se ha optado por la utilización de una interfaz de consola, este tipo de interfaz es la más habitual que podemos encontrar en traductores y compiladores. La línea de comandos es la interfaz que permite más flexibilidad en su uso, además en caso de querer automatizar las tareas, es muy fácil invocar a la aplicación dentro de un script, lo cual sería imposible con una interfaz gráfica.

La idea principal de interacción con el usuario es que una vez se haya descargado la aplicación, la añada al path y luego pueda utilizarla simplemente invocando perldoop desde la consola.

En la figura 7 podemos ver un ejemplo de la invocación de la aplicación, en este caso se ha optado por el parámetro –h, que despliega la documentación de aplicación mostrando los parámetros existentes y una breve descripción de lo que hacen.

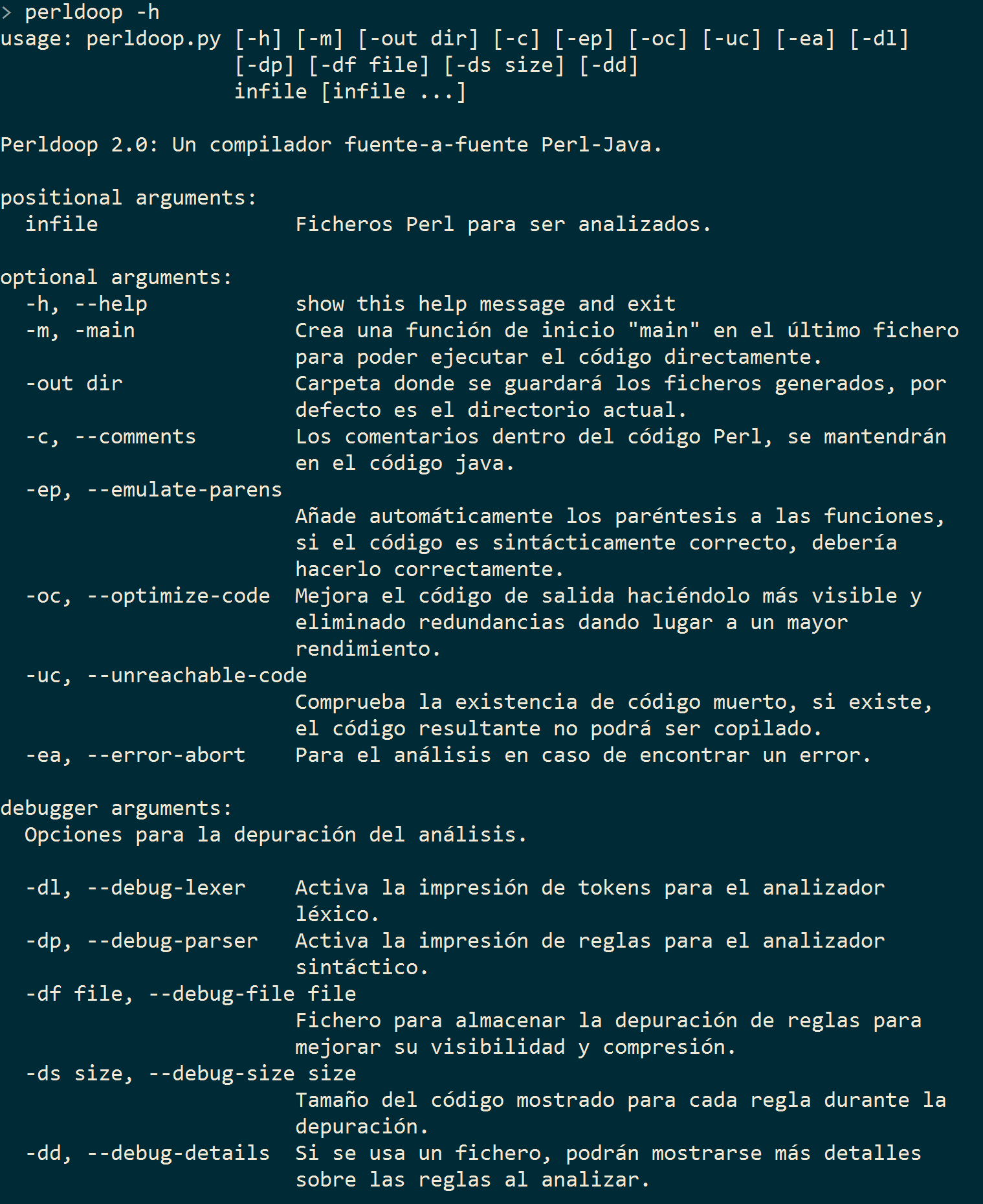


Figura 7: Ayuda Perldoop 2

## Analizador Léxico

El analizador léxico es la primera fase en un proceso de compilación o traducción, el analizador léxico recibe como entrada la secuencia de caracteres del código fuente y produce como salida una lista de tokens. Los tokens son un conjunto de caracteres que tienen un significado coherente dentro del lenguaje de programación en cuestión, por ejemplo ‘=’, ‘&&’,’while’ son tokens comunes que podemos encontrar.

El analizador léxico trabaja bajo demanda del analizador sintáctico devolviéndole un componente léxico según el analizador sintáctico lo va necesitando para avanzar en la gramática.

En este caso, el analizador sintáctico será el encargado de crear y pasarle el código fuente al analizador léxico.

## Analizador Sintáctico

El analizador sintáctico es el encargado de comprobar si una cadena de tokens(componentes léxicos) puede ser generada por una determinada gramática. En este caso se ha intentado definir una gramática lo suficientemente amplia para analizar la mayoría de los scripts usados en el CITIUS en el campo de procesamiento de lenguaje natural.

El ritmo del análisis lo lleva el analizador sintáctico, él es quien llama al analizador léxico y cuando tiene suficiente información llama al generador de código para ir creando la traducción. Una vez que el análisis ha terminado, la traducción ya ha sigo generada completamente y será retornada.

Normalmente los analizadores sintácticos son los encargados de generar un árbol sintáctico para luego verificar el código fuente. Para simplificar este proyecto y adaptarlo al tiempo requerido por un trabajo de fin de grado, se optado por fusionar el analizador sintáctico con el semántico y el generador de código dando como resultado un traductor con menos código y más rápido de implementar.

## Comprobación y generación de código

La comprobación y la generación de código, está formado por un conjunto de funciones agrupadas en módulos que son invocadas por el analizador sintáctico. La tarea de estas funciones es validar los parámetros recibidos del analizador sintáctico y si son correctos, generar una cadena con la traducción java correspondiente.

Las funciones reciben como parámetro todos los datos necesarios para realizar su tarea como por ejemplo la tabla de símbolos. La generación de código se basa en una simple concatenación de los argumentos y en algunos casos la conversión de un tipo de dato a otro.

## Sistema de errores

Cada vez que una fase del traductor falla, se invoca al sistema de error para informar al usuario de su naturaleza y donde se cometió. Los primeros errores que pueden ocurrir están a cargo de la interfaz de consola, estos corresponden a problemas con el sistema de archivos: fichero no existe, sin permiso de lectura, error de lectura, etc.

Lo primero que hace el sistema de error es invocar al sistema de mensaje para obtener el texto que se le mostrara al usuario, luego calculara la línea y la columna usando la posición del carácter y para terminar obtiene un fragmento del código fuente para mostrárselo al usuario. En la figura 8 podemos ver un ejemplo de varios errores de diferente tipo en el código fuente.

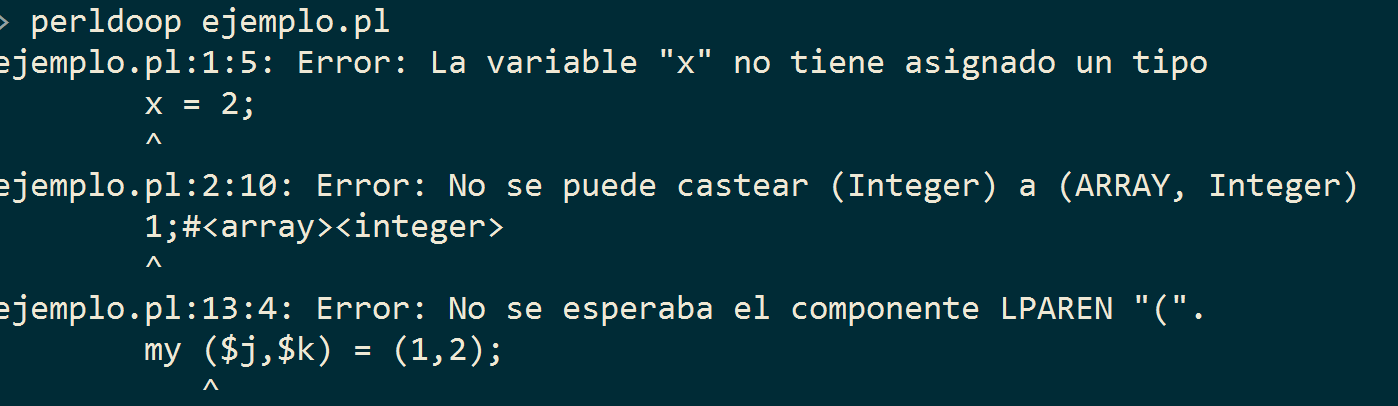


Figura 8: Ejemplo sistema de errores

## Sistema de mensajes

El sistema de mensajes contiene todos mensajes que son mostrados al usuario en algún momento, estos abarcan desde la ayuda de la interfaz de consola hasta los errores. Gracias a este sistema los mensajes no están mezclados con el código fuente y es fácil cambiarlos y adaptarlos según va evolucionando la aplicación.

Los mensajes están almacenados en una tabla en formato clave valor, donde la clave es un código usado para referenciar al mensaje dentro del código fuente. A su vez la tabla está asociada a un idioma dentro de una tabla de mayor nivel, de esta forma pueden crearse muchas tablas distintas en distintos idiomas. La elección del idioma debe cambiarse de manera manual en este fichero editando la constante LANGUAGE.

# Diseño e Implementación

Una vez definida la arquitectura del sistema, es el momento de detallar el diseño a seguir durante su la implementación. En este apartado de explicar en detalle el funcionamiento y las distintas clases del sistema, así como interactúan entre ellas.

## Diagrama de clases



Figura 9: Clases con sus atributos y métodos

La figura 9 muestra todas las clases del sistema, en ella podemos ver todos los atributos y métodos de los que está compuesta cada una, pero no sus relaciones. La mayoría de las clases están compuestas por muchos métodos simples lo cual hace que su representación en el diagrama sea muy grande, por esa razón en primer lugar se muestra la composición de las clases y luego en la Figura 10 se muestran las relaciones sin su contenido.

Siguiendo la filosofía de la programación extrema, el código fuente también se considera documentación y en caso de ser posible es mejor documentar el propio código a generar documentación externa. Por esa razón, todos los métodos y atributos mostrados en la Figura 9 están descritos en profundidad en el propio código y pueden ser consultados en el fichero .py correspondiente.

C:\Users\César\Desktop\diagramaclases.png

Figura 10: Diagrama de clases

Para simplificar el diagrama se ha agrupado las clases que contienen solo métodos estáticos y que nunca son instanciadas. A continuación, se hará un breve resumen de la finalidad y uso de cada clase:

* **Perldoop:** Clase que inicia todo el sistema, es la encargada de interpretar los argumentos y crear un analizador sintáctico por cada fichero de entrada.
* **Options:** Contiene las opciones pasadas como argumento a la aplicación y que el analizador sintáctico heredará. De esta forma tenemos un fichero que contiene las opciones por defecto y que un usuario podría ajustar sin perjudicar el resto de variables del Parser.
* **Parser:** Analizador sintáctico, tiene definida toda la gramática en su interior y las llamadas a las funciones de comprobación y generación.
* **Lexer:** Analizador léxico, tiene definidos los tokens y las palabras reservadas.
* **Declare:** Almacena la pre declaración de tipo de una variable hasta que esta es instanciada. La clase desaparece en cuanto la variable es declarada.
* **Messages:** Almacena todos los mensajes que se mostrar a la aplicación y es la encargada de imprimir los errores del sistema.
* **Code:** Almacena el código que es pasado en la reducción del analizador sintáctico, contiene todos los datos necesarios para decidir cuando existen varias traducciones posibles.
* **Access:** Almacena los accesos a variables hasta que la variable es usada, esto es necesario porque en caso de un HashMap no sabemos si habrá que realizar un get o un put hasta el último momento.
* **Position:** Guarda los datos relacionados con la posición dentro del fichero, estos datos son usados en caso de error para mostrarle fila y columna al usuario junto con una porción de código.
* **Variable:** Almacena una variable declarada, contiene los datos de tipo, nombre y acceso en el caso de que el fichero fuera marcado como paquete.
* **Funcion:** Contiene los datos de las funciones declaradas por el usuario dentro del código, aquí se almacenan los tipos de los argumentos y de sus retornos.
* **Package:** Una vez terminado el análisis sintáctico de un fichero, si está marcado como paquete, se almacenarán aquí los atributos y funciones para que puedan ser invocados en otro fichero.
* **Casting:** Contiene las funciones que permiten cambiar de tipo las expresiones en el código, los cambios se realizan mediante un casting directo o la invocación a las funciones java correspondientes.
* **Auxiliary:** Funciones varias que suelen ser usadas por todos los módulos, entre ellas se encuentran la identación de código o la creación de los imports.
* **Variables:** Tiene las funciones de gestión de variables de la tabla de símbolos y la generación de variables auxiliares reservadas para conversiones indirectas del código.
* **DataTypes:** Contiene las constantes de todos los datos y tipos, se usa una clase para no escribir las constantes como cadenas en las que podemos cometer errores.
* **Statements:** Genera el código de las sentencias Perl, declara las variables y valida si el código es necesario que sea traducido. Por ejemplo el código “1+1;” al no almacenar el resultado no generara ninguna traducción.
* **Functions:** Almacena la gramática de las funciones nativas de perl que el analizador sintáctico heredara. Las funciones de Perl tienen una sintaxis que entra en conflicto con el tipado, por eso deben ser definidas y tratadas aparte.
* **Blocks:** Genera el código de todos los bloques de código, desde los bloques de control a la clase que contiene toda la traducción.
* **Operations:** Genera el código de todas las operaciones que pueden realizarse con expresiones, ya se matemáticas, lógicas, comparaciones, etc.
* **Hadoop:** Contiene la gramática con las transformaciones de código para traducir un código secuencias por la estructura MapReduce paralela que usa Hadoop. Se realiza en una clase a parte para no condicionar el resto de la herramienta y permitir nuevas transformaciones en el futuro.
* **Collections:** Contiene el código para gestiones las colecciones, aquí es donde se inicializan y se verifica si los accesos son correctos en tipo.

## Diagramas de interacción

Los diagramas de interacción representan como se relaciona un usuario con el sistema y además como los objetos intervienen entre sí para realizar una función. En un traductor, las interacciones pueden ser muy largas y dependen mucho del código que estamos analizando así pues a continuación se mostrara el diagrama de interacción de un código simple que permita extrapolar el funcionamiento del sistema.

El código elegido será “my $x=2;#<string>” que declara una variable de tipo cadena y almacena en ella el valor numérico 2, para el cual tendrá que realizar una transformación de tipo.

C:\Users\César\Desktop\secuencia1.png

Figura 11: Diagrama de secuencia de un análisis

Como podemos ver en la Figura 11, un fragmento tan pequeño de código, ha creado un diagrama de secuencia considerablemente grande. El sistema sigue el mismo padrón para todos los tipos de código, primero va leyendo tokens y cuando tiene los suficientes datos para realizar una reducción, realiza una llamada al módulo de generación y comprobación de código y luego reduce. En la secuencia 27, vemos el comportamiento ante el final de sentencia, donde sus variables son declaradas y termina su procesado, todas las sentencias son luego concatenadas una vez que se crear el bloque que las contiene. En este caso al no existir ningún bloque son englobadas dentro de la clase en un método reservado para ello.

Un detalle es que una vez terminada la traducción y esta es retornada por la función parser, el main antes de escribirla en el fichero, procede a la identacion de código para hacerlo más visible al usuario. El usuario, no es notificado de ninguna manera durante el proceso de este análisis, igual que otros compiladores como ‘gcc’ solo se informará al usuario en caso de encontrar un error y si el usuario no recibe ningún mensaje quiere decir que la traducción se ha generado correctamente.

En el siguiente diagrama se muestra un fragmento de una iteración donde sucede un error porque se intenta leer una variable que no existe.

C:\Users\César\Desktop\secuencia2.png

Figura 12: Error en el análisis

En la figura 12 podemos el proceso seguido en caso de entrar un error. En primer lugar, debemos notificar al usuario que el código fuente contiene un error y luego avisar al analizador de que el código contiene errores. Una vez hecho esto, es necesario subsanar el error para poder continuar el análisis, en este caso crear la variable. El código javo ya no será impreso en un fichero al terminar, pero si queremos seguir el código en busca de más errores es necesario subsanarlo o puede producir una cascada de errores. Un ejemplo de esto es, una variable que no existe, nunca fue asignada, con lo cual si se intenta leer saldrán ambos errores si el primero no crea una variable libre de otros errores. El resto de errores siguen la misma filosofía y siempre se intenta que un mismo error nunca sea notificado de dos formas distintas.

## Diseño tokens analizador léxico

En este apartado definiremos los tokens que pueden ser reconocidos en el código fuente y el sistema para identificarlos. Según su origen podemos dividir los tokens en tres conjuntos diferentes: Básicos, Palabras reservadas y Etiquetas.

## Tokens Básicos

Los tokens básicos no todos tienen un valor fijo y son reconocidos por expresiones regulares en la mayoría de los casos, en la tabla 3 podemos ver todos los tokens definidos, la expresión regular que los reconoce y una pequeña descripción aclaratoria.

En algunos lenguajes de programación se puede definir los caracteres como tokens para simplificar su definición de forma que ‘+’ usa su valor ASCII como token. En Python al ser un lenguaje de muy alto nivel, es necesario definir una constante para todos los tokens.

|  |  |  |
| --- | --- | --- |
| Id | Expresión regular | Descripción |
| ID | [A-Za-z\_][\w\_]\* | Identificadores |
| DOLLAR | \$ | Escalar |
| AT | @ | Array |
| PERCENTAGE | % | HashMap/Modulo |
| INT\_NUMBER | (([1-9]\d\*)|  (o[xX][a-fA-F\d]+)|  (o[bB][01]+)|  (0[0-7]\*)) | Todos los números enteros en formato: Decimal, Hexadecimal, Octal y Binario |
| FLOAT | ((\d\*\.\d+)|(\d+\.\d\*)) | Números decimales |
| FLOAT\_NUMBER | (INT\_NUMBER | FLOAT)  [eE](\+-)?INT\_NUMBER | Números en notación científica. |
| STRING\_QUOTE | '([^'\n]|(\.))\*' | Cadenas simples |
| STRING\_DOUBLE\_QUOTE | “([^”\n]|(\.))\*” | Cadenas dobles |
| CMD | `.+` | Cadenas ejecutables |
| PLUS | \+ | Suma |
| MINUS | - | Resta |
| TIMES | \\* | Multiplicación |
| DIVIDE | / | División |
| POW | \\*\\* | Elevado |
| OR | | | Or binario |
| AND | & | And binario |
| NOT | ~ | Not binario |
| XOR | \^ | Xor binario |
| LSHIFT | << | Desplazar bit izquierda |
| RSHIFT | >> | Desplazar bit derecha |
| LOR | || | OR lógico |
| LAND | && | And lógico |
| LNOT | ¡ | Not lógico |
| NUM\_LT | < | Menor que |
| NUM\_GT | > | Mayor que |
| NUM\_LE | <= | Menor/Igual que |
| NUM\_GE | >= | Mayor/Igual que |
| NUM\_EQ | == | Igual que |
| NUM\_NE | ¡= | No igual que |
| STR\_REX | =~ | Reconocido por regex |
| STR\_NO\_REX | !~ | No reconocido por regex |
| SMART\_EQ | ~~ | Comparador Inteligente |
| CMP\_NUM | <=> | Comparador números |
| EQUALS | = | Asignación |
| TIMESEQUAL | \\*= | Asignar multiplicación |
| DIVEQUAL | /= | Asignar división |
| MODEQUAL | %= | Asignar modulo |
| POWEQUAL | \\*\\*= | Asignar elevado |
| PLUSEQUAL | \+= | Asignar suma |
| MINUSEQUAL | -= | Asignar resta |
| LSHIFTEQUAL | <<= | Asignar desp izquierda |
| RSHIFTEQUAL | >>= | Asignar desp derecha |
| ANDEQUAL | &= | Asignar And binario |
| OREQUAL | |= | Asignar Or binario |
| XOREQUAL | \^= | Asignar Xor binario |
| LANDEQUAL | &&= | Asignar And lógico |
| LOREQUAL | ||= | Asignar Or lógico |
| PERIODEQUAL | \.= | Asignar Concatenación |
| XEQUAL | x= | Asignar Repetición |
| PLUSPLUS | \+\+ | Incrementar |
| MINUSMINUS | -- | Decremento |
| LPAREN | \( | Paréntesis derecho |
| RPAREN | \) | Paréntesis izquierdo |
| LBRACKET | \[ | Corchete derecho |
| RBRACKET | \] | Corchete izquierdo |
| COMMA | (, | =>) | Los dos tipos de comas |
| PERIOD | \. | Concatenación |
| TWO\_PERIOD | \.\. | Generador de rango |
| QUEST\_CLOSE | \? | Operador ternario |
| SEMI | ; | Punto y coma |
| COLON | : | Else del Operador ternario |
| TWO\_COLON | :: | Acceso a paquete |
| LBRACE | \{ | Llave izquierda |
| RBRACE | \} | Llave derecha |
| BACKSLASH | \\ | Barra invertida/Escape |
| POINTED | -> | Acceso a puntero |
| REGEX\_MOD | (i|s|m|x|o|p|d|a|  u|l|g|cg|e)? | Modificador expresiones regulares |
| M\_REGEX | m/.\*/ REGEX\_MOD | Regex tipo match |
| S\_REGEX | s/.\*/.\*/REGEX\_MOD | Tegex tipo sustitución |
| Y\_REGEX | (y|tr)/.\*/.\*/REGEX\_MOD | Regex tipo transposición |
| STDIN | <STDIN> | Entrada estándar |
| COMMENT | \# | Inicio comentario |
| COMMENT\_SPECIAL | \#!.\* | Comentarios especiales |
| COMMENT\_IGNORE\_LINE | .\*\#<ignore-line> | Ignorar la línea |
| COMMENT\_IGNORE\_BLOCK | \#<ignore-block>\n  (.|\n)  \*\#<ignore-block> | Ignorar bloque de código |
| COMMENT\_CODE\_IMPORT | \#<java-import>.\* | Añadir import java |
| COMMENT\_CODE\_LINE | \#<java-line>.\* | Añadir sentencia java |

Tabla 3: Tokens básicos

## Tokens de palabras reservadas

Las palabras reservadas son identificadoras que el lenguaje de programación Perl tiene reservado para un uso especial. Para diferéncialos, cada vez que el analizador encuentra un identificador, consulta un hashMap que contiene todas las palabras reservadas y le asigna el tipo de la misma. En Perl todas las variables son precedidas por $,@,% por lo tanto nunca son variables reservadas, por otro lado, la funciones no tienen ningún símbolo precedido con lo cual son las que deben diferenciarse de las palabras reservadas.

La tabla 4 contiene todas las palabras reservadas definidas en el analizador, el valor de la palabra reservada corresponde al id en minúsculas a no ser que se especifique lo contrario.

|  |  |
| --- | --- |
| Id | Descripción |
| MY | Crea variables |
| SUB | Define métodos |
| OUR | Crea variables accesibles en paquete |
| PACKAGE | Crea un paquete |
| WHILE | Bucle while |
| DO | Bucle do while |
| FOR (for,foreach) | Bucle for y for each, por convención de Perl se usan indistintamente. |
| UNTIL | Condicional if negada |
| IF | Condicional if |
| ELSIF | Contracción de else if |
| ELSE | Bloque else |
| UNLESS | Bucle while negado |
| LAST | Rompe el bucle |
| NEXT | Sala a la próxima iteración del bucle |
| CMP | Compara dos cadenas |
| X | Repite una cadena un número de veces |
| UNDEF | Valor indefinido, similar a null java |
| LLAND (and) | And lógico de baja precedencia |
| LLOR (or) | Or lógico de baja precedencia |
| LLXOR (xor) | Xor lógico de baja precedencia |
| LLNOT (not) | Not lógico de baja precedencia |
| STR\_LT (lt) | Menor que para cadenas |
| STR\_GT (gt) | Mayor que para cadenas |
| STR\_LE (le) | Menor/Igual que para cadenas |
| STR\_GE (ge) | Mayor/igual que para cadenas |
| STR\_EQ (eq) | Igual que para cadenas |
| RSTR\_NE (ne) | No igual que para cadenas |

Tabla 4: Tokens de palabras reservadas

A parte de las palabras reservadas a las que nos tienen acostumbrados todos los lenguajes de programación, Perl nos añade operadores de baja precedencia (para evitar entre paréntesis sus argumentos) y operadores de comparación de cadenas.

## Tokens de Etiquetas

Las etiquetas son el punto clave para el funcionamiento de Perldoop, con ellas podemos traducir un lenguaje sin tipo a un lenguaje con tipo y realizar diversas transformaciones en el código.

Las etiquetas se definen dentro de comentarios para no afectar al correcto funcionamiento del script con su interprete nativo. Las etiquetas se definen como caracteres alfanuméricos encerrados entre ‘<’ y ‘>’. Como vimos en el apartado de tokens básicos, el comentario eta definido solo por ‘#’, al reconocer este carácter el analizador léxico entra en un estado para reconocer etiquetas. Si se encuentra un elemento con sintaxis de etiqueta, se procederá a comprobar si existe, en caso afirmativo se devolverá como token, en caso contrario se añadirá al comentario.

En la tabla 5 podemos ver los patrones de las etiquetas existentes.

|  |  |  |
| --- | --- | --- |
| Id | Expresión regular | Descripción |
| LABEL\_ID | < [A-Za-z\_][\w\_]\*> | Identificador de etiquetas |
| LABEL\_SIZE | <[1-9][0-9]\*> | Etiquetas de tamaño |
| LABEL\_VAR | <(\$|@|%) [A-Za-z\_][\w\_]\*> | Etiquetas de variables |

Tabla 5: Exprsiones regualres para etiquetas

Las etiquetas de tamaño y variables sirven para hacer referencia a variables en predeclaracion y para definir tamaños en caso de colecciones. Los identificadores sin embargo siguen la filosofía del reconocimiento de palabras reservas, se comprueba y solo se devuelve si existe. A continuación, se muestran las etiquetas que son reconocidas por estas 3 expresiones regulares:

|  |  |
| --- | --- |
| Id | Descripción |
| VAR | Todas las variables reconocidas con LAVEL\_VAR |
| SIZE | Todos los tamaños reconocidos con LABEL\_SIZE |
| L\_ARRAY | Identificador del tipo de dato Array |
| L\_HASH | Identificador del tipo de dato Hash |
| L\_LIST | Identificador del tipo de dato List |
| REF | Identificador del tipo de dato Referencia |
| TYPE | Identificador de los tipos de dato básicos |
| ARGS | Identificador de los argumentos de una función. |
| RETURNS | Identificador de los retornos de una funcion |
| MAPPER\_CODE | Identificador de un Mapper |
| MAPPER\_LOOP | Identificador del bucle a paralelizar en un Mapper |
| HADOOP\_PRINT | Identificador de las impresiones en Hadoop |
| REDUCER\_CODE | Identificador de un Reducer |
| REDUCER\_OP | Identificador de la operación de Reduccion |
| REDUCER\_CHANGE | Identificador de la operación de cambio de clave en Reducer. |
| REDUCER\_KEY | Identificador de la variable clave en Reducer |
| REDUCER\_VALUE | Identificador de la variable valor en Reducer |
| REDUCER\_VAR | Identificador de las variables a declarar en Reducer |

Tabla 6: Tokens de etiquetas

## 8.4 Gramática Analizador Sintáctico

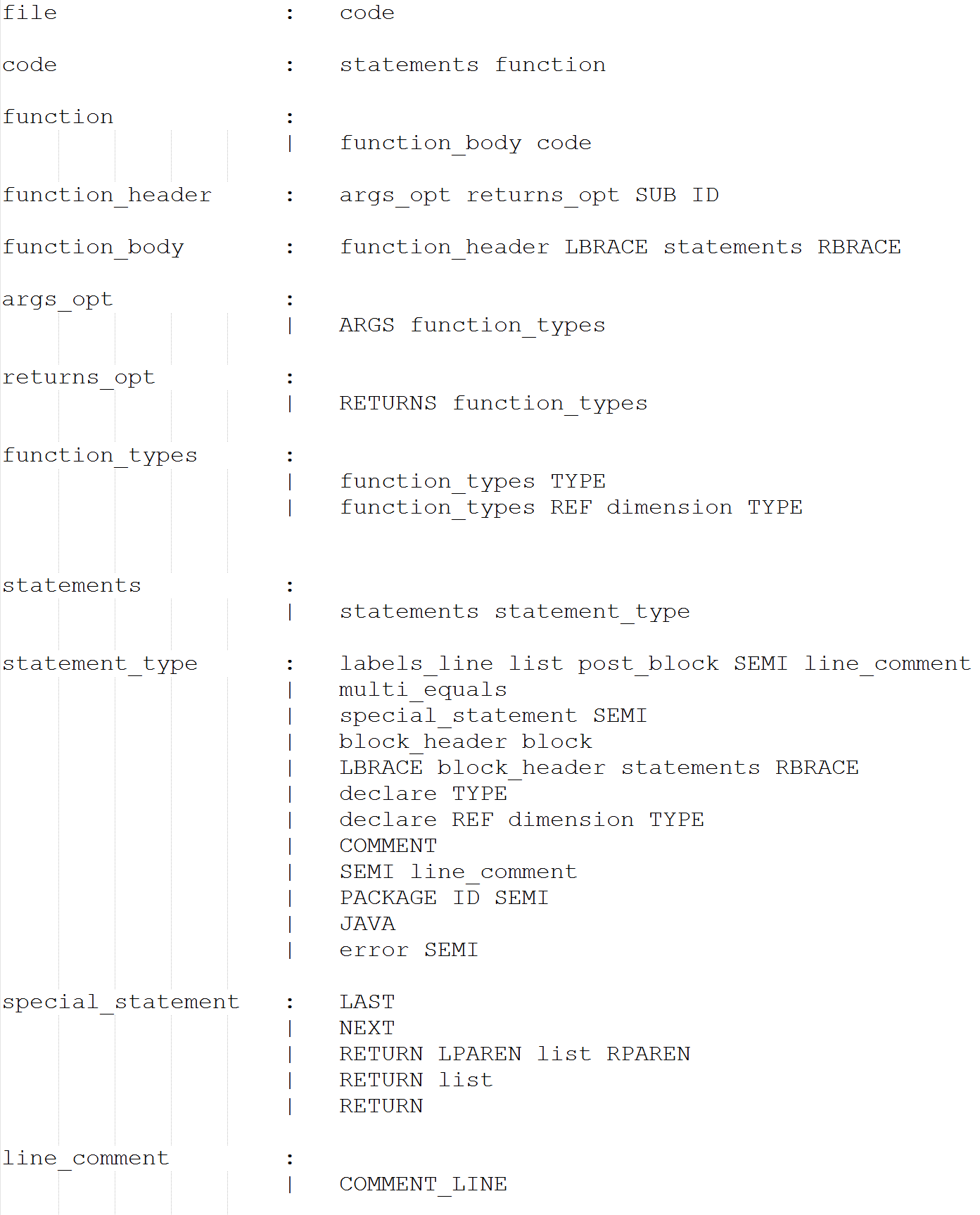
El analizador sintáctico es el encargado de la validación de los tokens para comprobar si la sintaxis es correcta. Existen diversos tipos de analizadores, ascendentes y descendentes, así como distintos tipos de gramáticas para cada de ellos que proporcionan las capacidades para definir un comportamiento determinado.

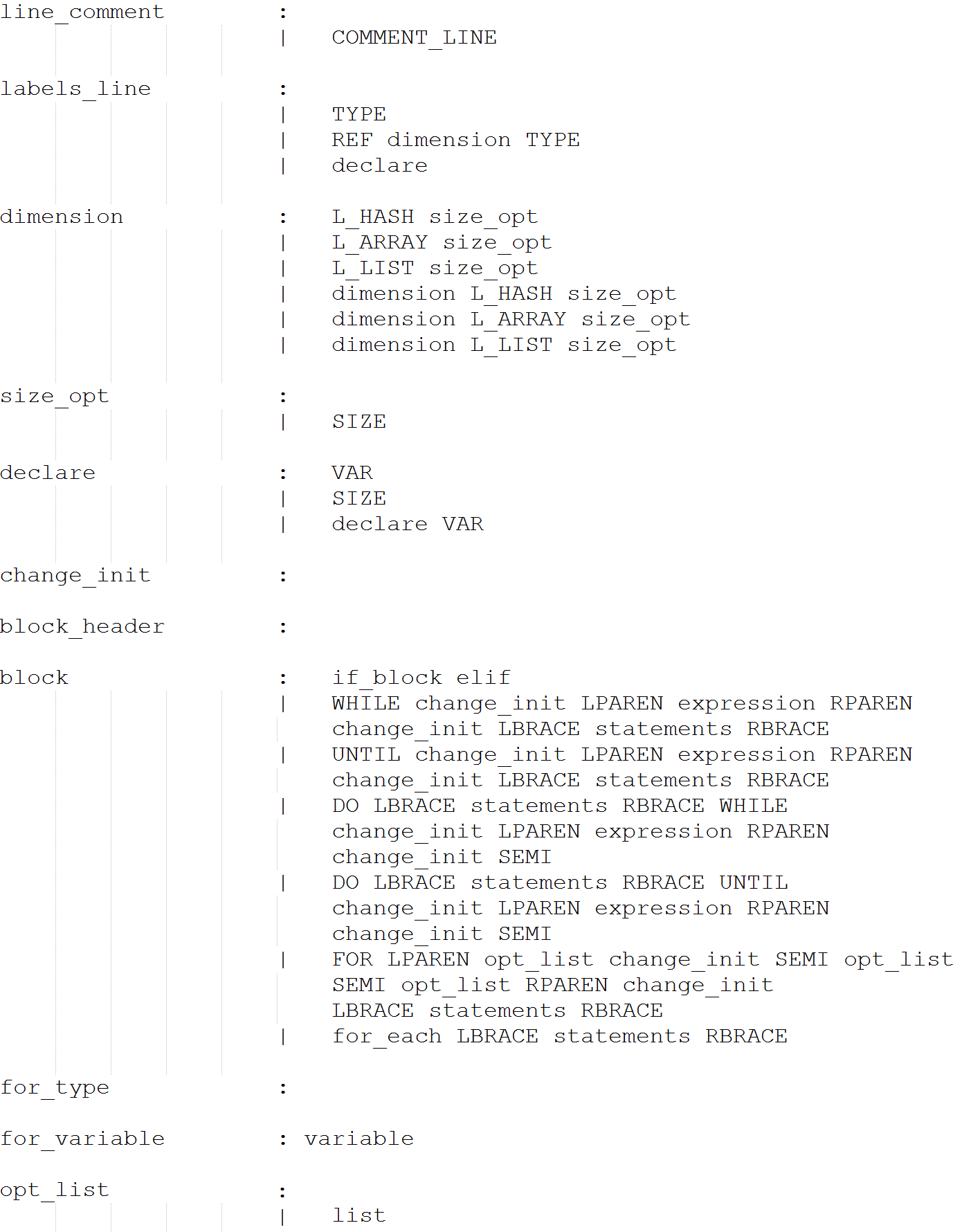
Este traductor hace uso de un analizador ascendente con una gramática LALR(1), esta gramática permite definir analizadores rápidos y que no consumen tanta memoria como los LR.

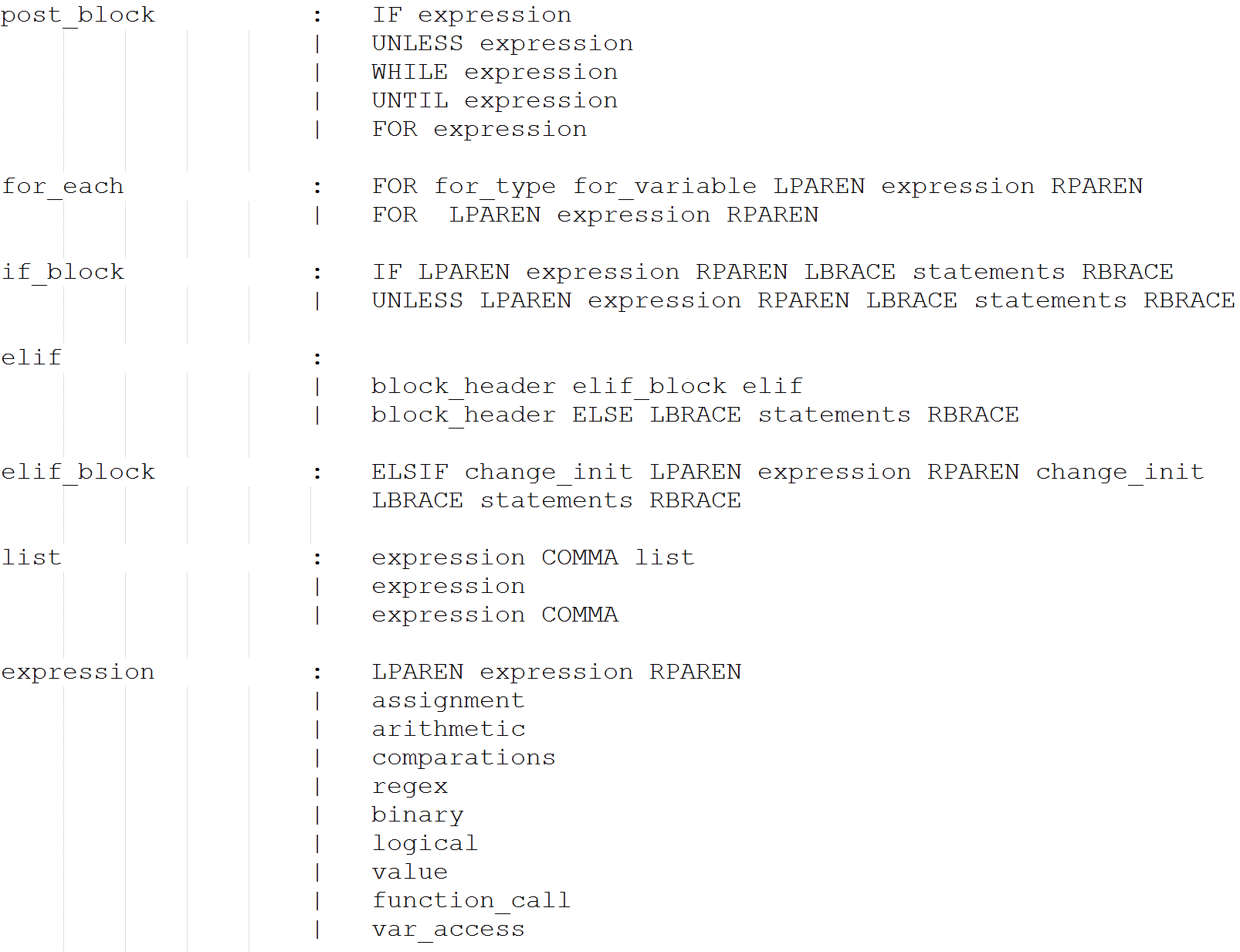
Perl tiene definido una sintaxis increíblemente grande y según su documentación, no pude ser reconocido por un analizador sintáctico, pero para problemas acotados como son los scripts de procesamiento de texto, una gramática LALR(1) es más que suficiente para realizar esta tarea.

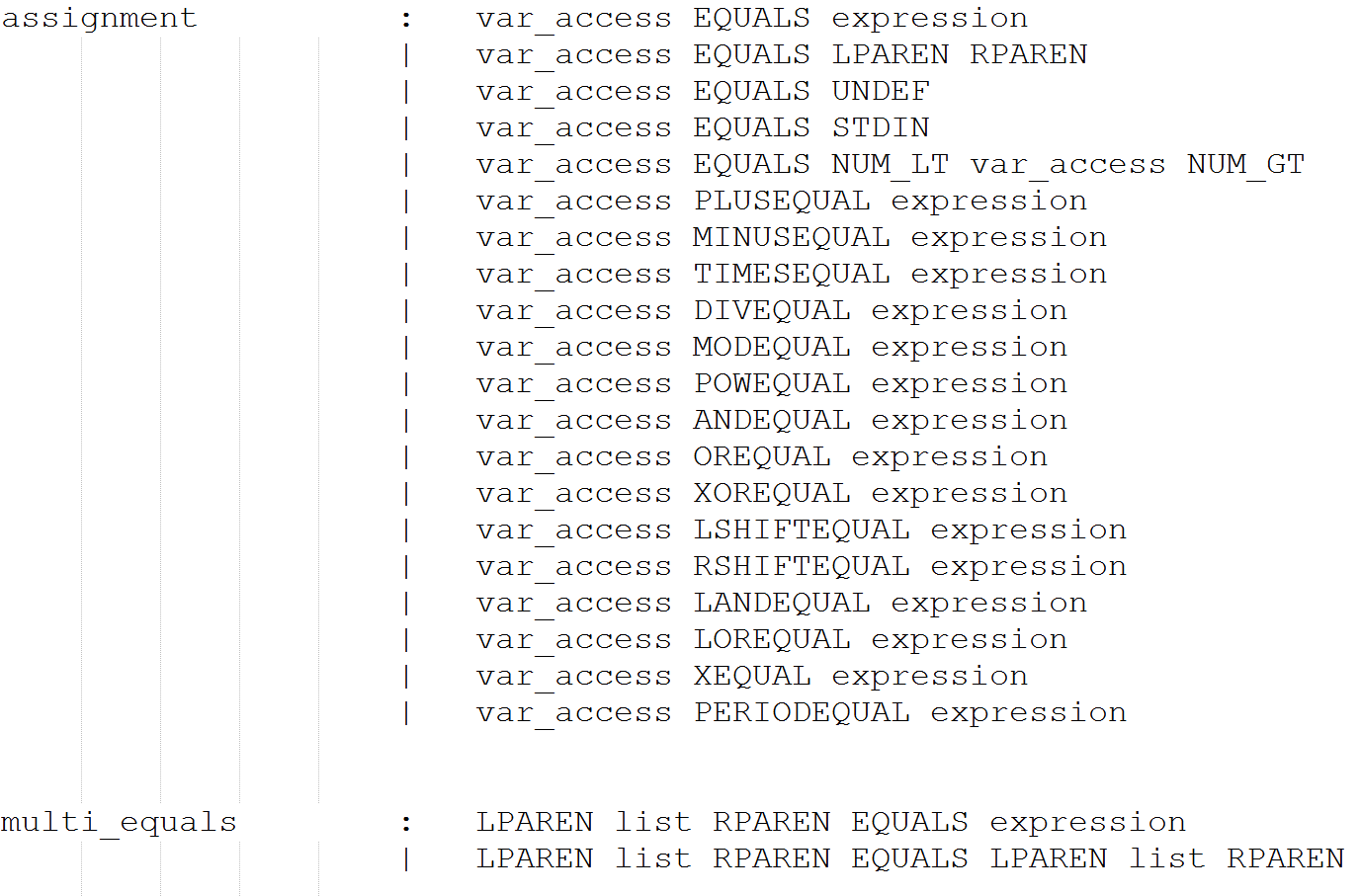
En la Figura 13 se representa la gramática definida en el analizador léxico, los componentes termianles de la gramatica se representan en mayúsculas y correspontes a los tokens definidos en el apartado de definición de analizador léxico. Por otro lado, los símbolos no terminales de la gramática se estriben en minúscula y corresponden a la parte izquierda de las reglas gramaticales.

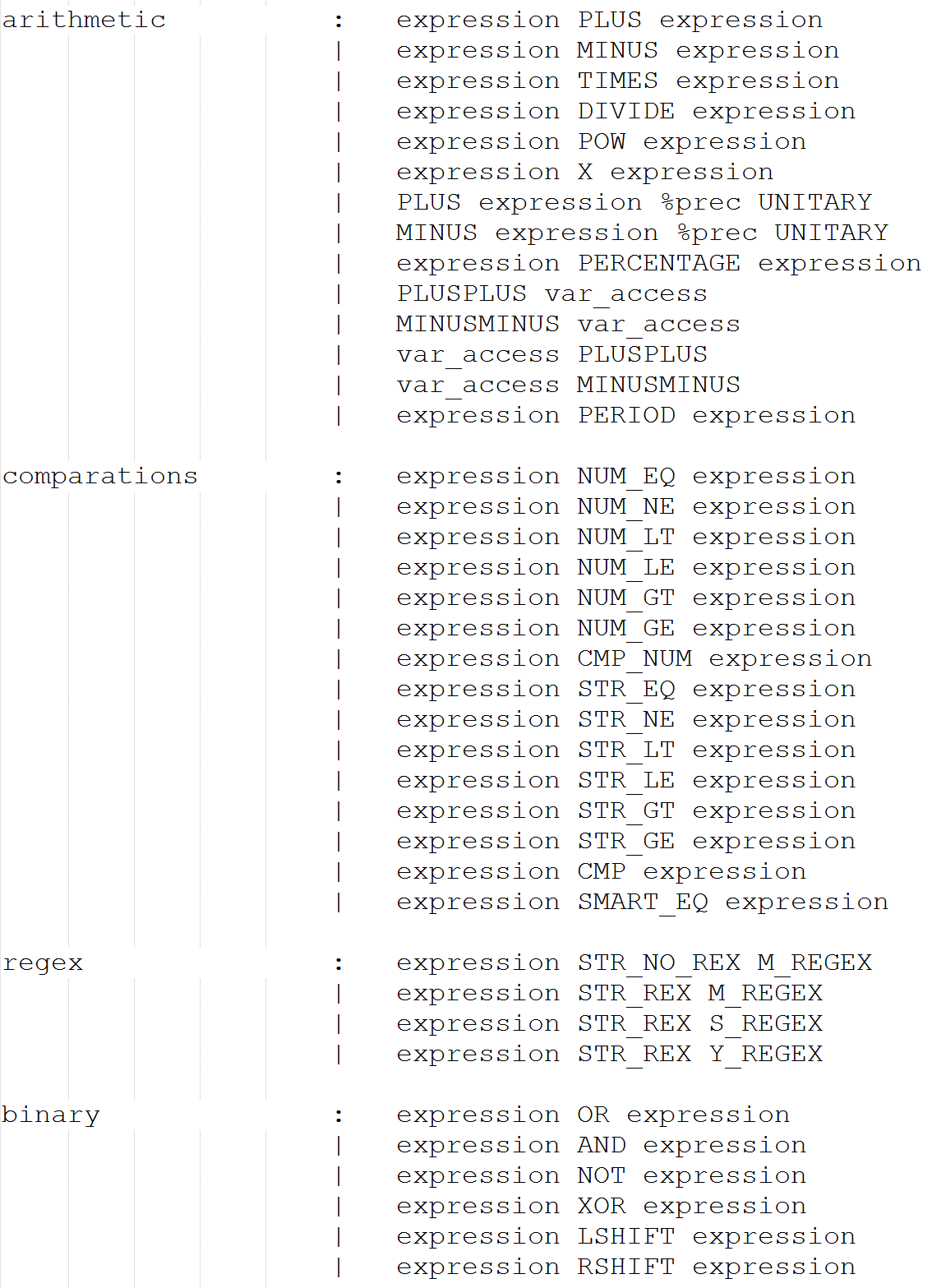
Figura 13: Gramática analizador sintáctico

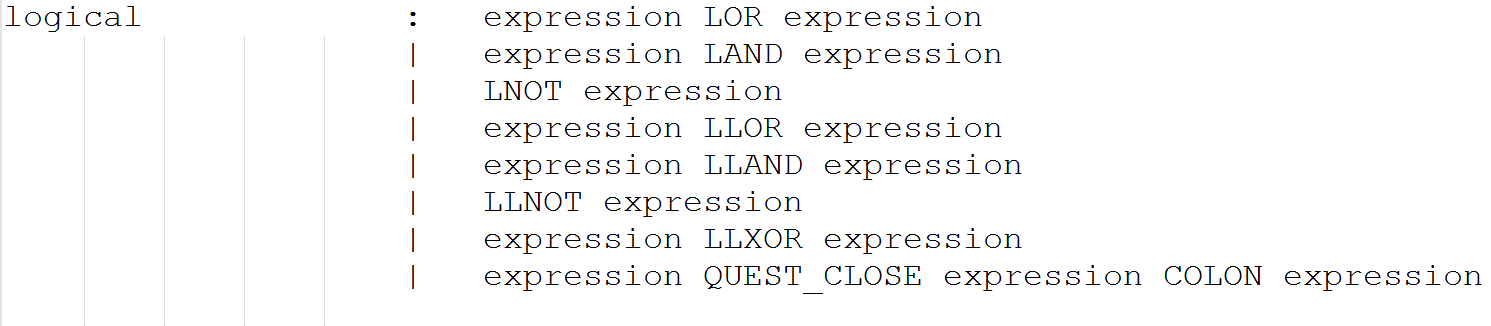


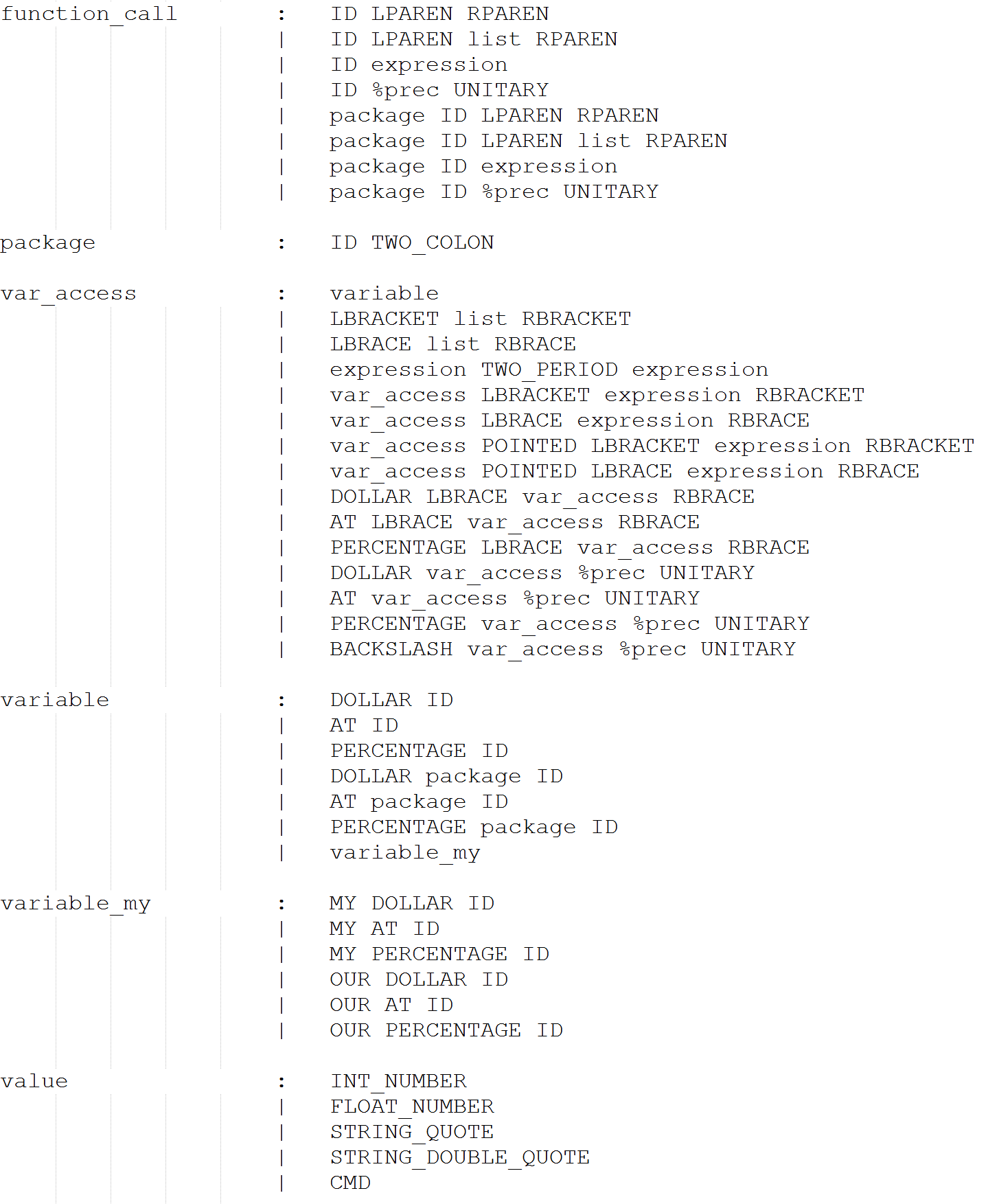












Para el funcionamiento de la gramática anterior, es necesario establecer una asociatividad y una precedencia para poder eliminar los conflictos en el análisis.

La asociatividad indica porque lado se empieza a reducir una regla, existen tres tipos, derecha, izquierda o nula. Por ejemplo, la operación ‘2+3+1’ pude realizarse de dos maneras, por la derecha (2+4) o por la izquierda (5+1) en este caso el resultado sería el mismo, pero aun así hay que establece un criterio firme a realizar.

La precedencia indica la prioridad entre operadores, igual que el caso anterior existen tres tipos disjuntos de precedencias, mayor, menor e igual. Cuando dos operadores se encuentran juntos deberá decirle prioridad uno al otro y en caso de tener igual prioridad se hará uso de la asociatividad para resolver la situación. Por ejemplo, en las matemáticas las multiplicaciones tienen prioridad frente a las sumas.

En la figura 14 podemos ver la asociatividad de las operaciones adaptada de la documentación oficial de Perl [16] definida en la sintaxis de PLY.



Las precedencias están ordenadas de menos presente a mas precedente, de forma que cuanto más abajo se encuentre un símbolo, más precedencia tendrá.

## Generación y Comprobación de código

En este apartado hablaremos sobre las particularidades a alto nivel de la comprobación y generación de código.

## Tipos de datos

Los tipos de datos son atributos que indican a java la clase de dato que va a usar, para obtener este tipo de dato se usan las etiquetas presentes en el código cuando se trata de variables o funciones, o es deducido en caso de las contantes.

## Tipos asociados a etiquetas

|  |  |
| --- | --- |
| Etiqueta | Tipo de dato java |
| <boolean> | Boolean |
| <interger> | Integer |
| <long> | Long |
| <float> | Float |
| <double> | Double |
| <string> | String |
| <file> | PerlFile |
| <array> | [] |
| <list> | List<> |
| <hash> | Map<> |
| <ref> | Ref<> |

Tabla 7: Relación etiqueta tipo de dato

Las etiquetas de tipos de dato básico fueron diseñadas con los nombres de los tipos de datos java para una traducción más cercana al usuario. Estos son los tipos de datos asociados a los escalares de Perl.

Los arrays de Perl pueden traducirse por Arrays o List java, la razón de realizar esta diferenciación es mejorar la eficiencia de acceso a cada uno de ellos. Por ejemplo, si se necesita un array estático que siempre tendrá un tamaño fijo, se deberá usar la etiqueta <array>, pero si no se conoce el tamaño o se van usar funciones para añadir elementos como push o pop, es las rápido usar <list>. A parte de esto ambas pueden usarse indistintamente siendo la única penalización el rendimiento.

Para simular las referencias, se ha creado una clase Ref que almacena una colección simulando el mismo comportamiento. La razón de que solo se pueden usar colecciones es que los tipos de dato básico son inmutables.

Los ficheros son unificados con una clase llamada PerlFile que permite tanto abrir ficheros para lectura como para escritura, de esta forma no es necesario especificar varios tipos de variables para ficheros.

## Tipos asociados a contantes

Los tipos de datos asociados a constantes, se evalúan de la misma forma que lo haría java si estuviese escrito en su propio código.

Los números sin decimales son siempre interpretados por Integer, estos pueden estar escrito en cualquier formato, Binario, Octal, Decimal o hexadecimal. Por otro lado, los números decimales y en notación científica son por defecto Double.

Las cadenas en perl pueden especificarse de tres formas diferentes:

* Comillas simples (‘’): Las constantes entre comillas simples son traducidas directamente cambiándolas por comillas dobles puesto que la semántica es la misma.
* Comillas dobles (“”): Las comillas dobles indican que en Perl la constate puede contener variables que deberán ser interpoladas, pero como eso no está soportado, la tradición se hace sin examinar la cadena.
* Comillas inclinadas (``): Las comillas inclinadas indican que la cadena deberá ser ejecutada por el sistema y luego almacenar el resultado. La traducción realiza el mismo comportamiento generando una función que retorna un String con el resultado de la llamada al sistema.

## Casting automático

El traductor cuenta con un sistema de casting automático para ayudar a lidiar con los problemas que aparecen al asignar tipos al código fuente.

Por defecto todos los tipos básicos pueden ser casteados directamente entre sí de igual forma que lo hace Perl. Esto no quiere decir que siempre sea semánticamente correcto, pero esa parte recae en el usuario.

Con las colecciones, el casting no se podrá hacer directamente igualando como los tipos básicos, el programador deberá recorrer la colección y copiar elemento a elemento. La única excepción a esta regla es entre un Array y una Lista del mismo tipo, dado que esencialmente son el mismo y debe fomentarse el uso entre ellos por factores de rendimiento.

A continuación, se muestran varias tablas con la forma de transformar para cada tipo de dato, las tablas solo incluyen las transformaciones validas, las obviadas darán error.

|  |  |
| --- | --- |
| Integer | |
| Boolean | ?1:0 |
| Long | (int) |
| Float | (int) |
| Double | (int) |
| String | Integer.parseInt |
| File | 1 |
| Array | length |
| List | size() |
| Map | size() |
| Ref | 1 |

|  |  |
| --- | --- |
| Long | |
| Boolean | ?1l:0l |
| Integer | = |
| Float | (long) |
| Double | (long) |
| String | Long.parseLong |
| File | 1l |
| Array | length |
| List | size() |
| Map | size() |
| Ref | 1l |

|  |  |
| --- | --- |
| Boolean | |
| Integer | != 0 |
| Long | != 0 |
| Float | != 0 |
| Double | != 0 |
| String | != “” && != “0” |
| File | true |
| Array | length != 0 |
| List | size() != 0 |
| Map | size() != 0 |
| Ref | != null |

|  |  |
| --- | --- |
| Double | |
| Boolean | ?1d:0d |
| Integer | = |
| Long | = |
| Float | = |
| String | Double.parseDouble |
| File | 1d |
| Array | length |
| List | size() |
| Map | size() |
| Ref | 1d |

|  |  |
| --- | --- |
| String | |
| Boolean | String.valueOf |
| Integer | String.valueOf |
| Long | String.valueOf |
| Float | String.valueOf |
| Double | String.valueOf |
| File | “1” |
| Array | String.valueOf(  length |
| List | String.valueOf(  size() |
| Map | String.valueOf(  size() |
| Ref | “1” |

|  |  |
| --- | --- |
| Float | |
| Boolean | ?1f:0f |
| Integer | = |
| Long | = |
| Double | (float) |
| String | Float.parseFloat |
| File | 1f |
| Array | length |
| List | size() |
| Map | size() |
| Ref | 1f |

|  |  |
| --- | --- |
| List | |
| Array | toList |

|  |  |
| --- | --- |
| Array | |
| List | toArray |

En algunos casos, se ha introducido valores constantes debido a la imposibilidad de aplicar el casting. Por ejemplo, si casteamos una referencia a un entero, deberíamos obtener la dirección de memoria donde se almacena el valor, con lo cual su valor no es importante.

## Definición de funciones

En este apartado se discutirá la transformación de las funciones Perl a funciones java, los principales problemas y las soluciones encontradas.

En Perl las funciones reciben sus argumentos en forma de array, el resultado de ese array es la concatenación de todos los elementos pasados como argumentos. El primer problema reside cuando uno de los argumentos es también un array, a continuación, se ilustra el comportamiento del mismo.

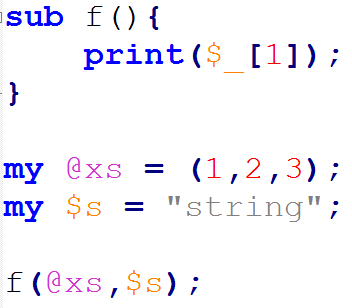


Figura 14: Función perl con array y escalar

El comportamiento esperado por una persona que no conoce el lenguaje Perl es pensar que la función imprimirá la cadena “string” pero en realidad se imprimirá el número 2. La razón de este comportamiento es que perl concatena la lista con escalar creando un nuevo array con cuatro argumentos. En java no podemos permitir este comportamiento debido a que, al no conocer la longitud de los arrays en tiempo de compilación, es imposible saber dónde empieza y donde acaba un array, lo cual impide realizar el casting al tipo correspondiente.

Para solucionar este problema, es necesario que Perl no concatene los argumentos y que si la función recibe dos argumentos, la función reciba un array de dos elementos. La solución pasa por obligar al usuario a pasar todas las colecciones por referencia en vez de por valor, de eso modo perl no concatenara y el usuario puede des referenciar luego en la función.

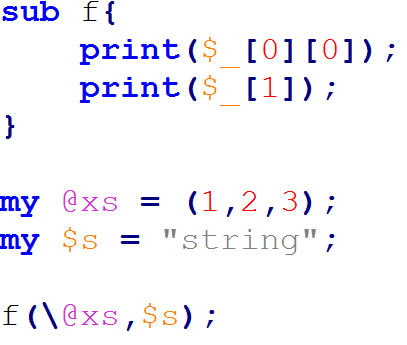


Figura 15: Funcion Perl con referencia Array y escalar

En la figura 15 podemos ver cómo podemos el array y el escalar se mantienen separados en la función, el primer elemento es un array al que podemos acceder a sus posiciones y el segundo es un escalar. Con esto el problema de las funciones queda solucionado y lo único que obligamos a hacer al usuario es añadir un ‘\’ antes de las colecciones en la llamada a las funciones.

Ahora falta abordar el tipado de las funciones tanto para los argumentos como para el retorno. Para ello se hará uso del mismo método que para el tipado de variables, a continuación, se muestra la sintaxis con varios ejemplos de funciones y su traducción.

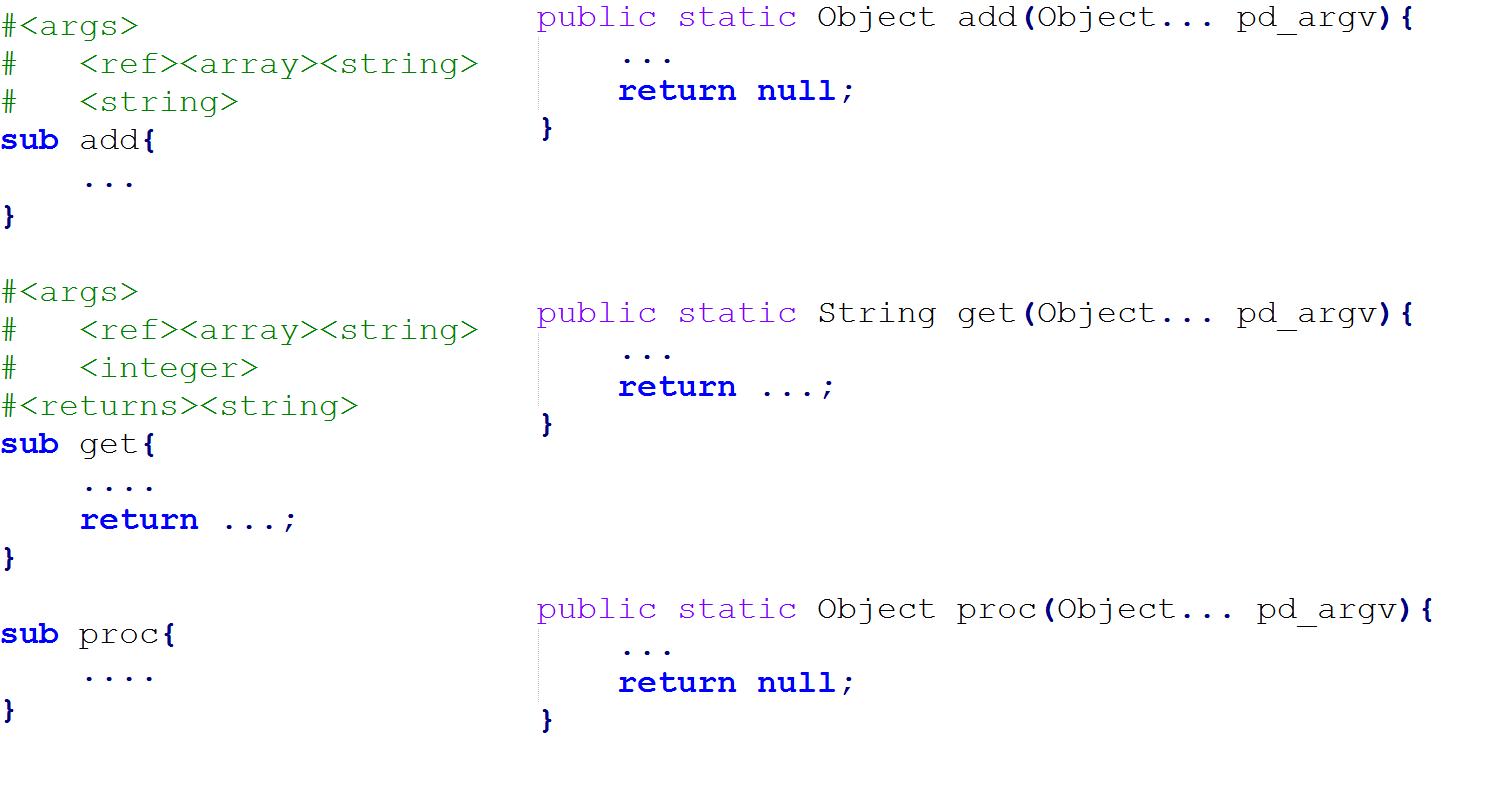


Figura 16: Ejemplo de tipado de funciones y su traducción

Como podemos ver en primer lugar el tipado de las funciones se realiza encadenando los tipos uno detrás de otro predecidos <args> para los argumentos y <returns> para los retornos. La posición de las etiquetas en varias líneas es meramente visual, puede optarse por poner todo en la misma línea.

Por otro lado, la traducción de las funciones es prácticamente igual para las tres funciones, todas las funciones reciben un array de Object donde estarán los argumentos en caso de que existan. Cuando el usuario obtiene un argumento se le aplica el cast en función de las etiquetas de la cabecera. Los retornos siguen un procedimiento parecido, si una función retorno un elemento, se le asignará ese tipo al retorno y si no retorna ninguno se conservará object por razones técnicas relativas al anidamiento de funciones.

Como nota final, Perl permite retornar más de un elemento en una función, pero ese caso es el mismo que para los argumentos, se retorna un array de object y se usa la cabecera de la función para aplicar el casting en la asignación.

## Asignación

En Perl, el operador igual ‘=’ puede usarse en cuatro contextos distintos, en este apartado se explicará el diseño para cada una de sus traducciones.

**Expresión = Expresión**

El primer caso es el contemplado por todos los lenguajes de programación, una expresión es almacenada en una variable. Por esa razón la traducción de este tipo es directa.

**Lista = Lista**

En el segundo caso, una lista de expresiones es asignada atómicamente a una lista de variables, por esa razón puede usarse para intercambiar variables y realizar acciones en grupo.

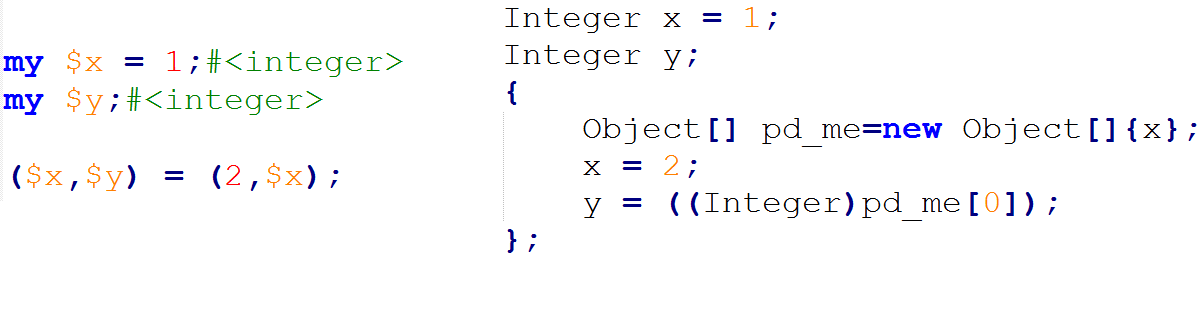


Figura 17: Asignación entre listas

Como podemos ver en la figura 17, queremos asignar el valor de un número y una variable a dos variables. La estrategia para realizar esta asignación es la siguiente:

* Abrimos un bloque para no conservar la variable auxiliar pd\_me y poder reutilizarla.
* Guardamos todas las variables del lado derecho en un array.
* Generamos una lista de asignaciones donde las constantes se copian directamente y las variables se sacan del array devolviéndoles su tipo.

Siguiendo estos pasos, las asignaciones se realizan de la misma forma que se haría en Perl conservando la atomicidad.

**Lista = Expresión**

El tercer caso es el usado para asignar a una lista de variables los elementos consecutivos almacenados en un array.

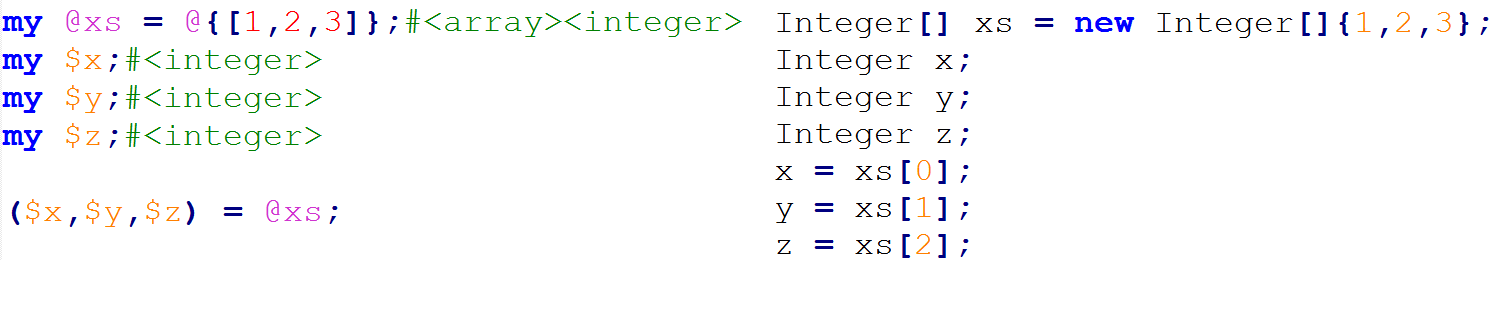


Figura 18: Asignación entre lista y expresión

Como podemos ver en la figura 18, tenemos un array de tres elementos que queremos asignar a tres variables distintas. La traducción es este caso es bastante sencilla, se va accediendo a las posiciones del array y luego se va realizando la asignación.

Pero qué pasaría si el array no estuviera almacenado en una variable como en este caso, por ejemplo, en el retorno de una función.

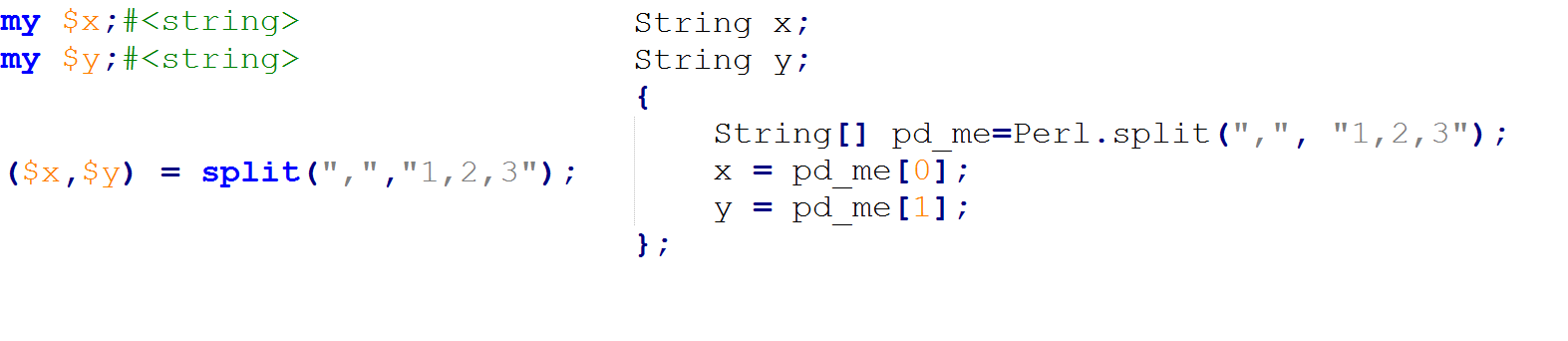


Figura 19: Asignación entre lista y función

La figura 19 muestra el comportamiento de en caso de no poder asegurar que el array sea una variable, se crea una variable auxiliar para realizar la asignación.

**Expresión = Lista**

El cuarto caso es el más complicado y el que no está soportado directamente, las listas en perl pueden indicar la inicialización de un array o de un hash por esa razón no se conoce el tipo hasta su asignación. Por esa razón se optado al igual que en las funciones por hacer uso de los punteros. En perl los punteros a array se inicializan enter [] y los punteros a hash se inicializan entre {}. Por esa razón las inicializaciones se realizarán creando un puntero y accediendo a él.

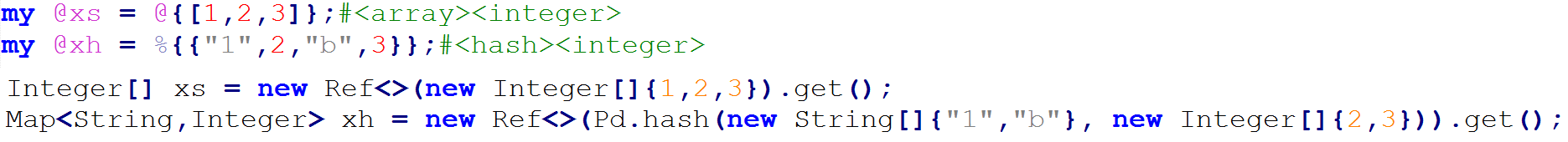


Figura 20: Inicialización de colecciones

Como podemos ver en la figura 20, la traducción parece un poco ineficiente, pero este es uno de los casos contemplados en las optimizaciones del traductor (Acceso a una referencia recién creada).

Si ejecutamos el traductor con la opción de optimizar el resultado será el siguiente.

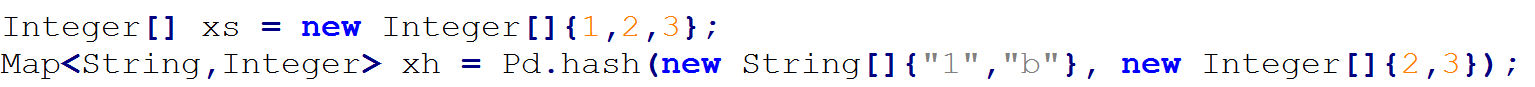


Figura 21: Inicialización de colecciones optimizada

Como podemos ver las referencias en el código java han sido eliminadas y el código es igual de eficiente que si hubiesen sido declaradas directamente.

## Ficheros

La gestión de ficheros es un asunto importante en los scripts de procesamiento de lenguaje natural. En Perl existen diferentes formas de gestionar los ficheros, pero para simplificar las traducciones se ha decido optar por el más usado. Las traducciones hacen uso de la librería java que se explicaría mas adelante pero aun así no hace falta para entender las traducciones.

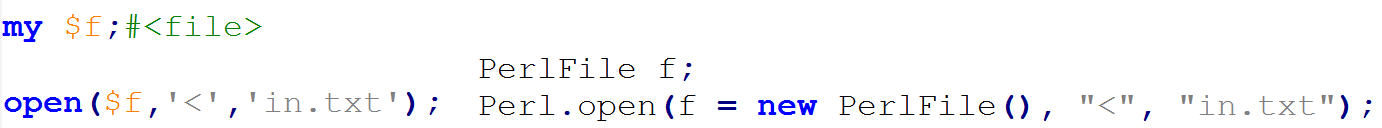


Figura 22: Apertura de fichero

Como podemos ver en la figura 22, la traducción se parece mucho al código nativo Perl, la variable se inicializa en el argumento de la función y es java en tiempo de ejecución quien realiza el resto del proceso.

En este caso el fichero se abriría para lectura por el uso del modificador ‘<’, los otros modificadores disponibles son ‘>’ y ‘>>’ para abrir un fichero para escritura para truncar y concatenar respectivamente.

La lectura de un fichero se puede realizar de dos maneras, línea a línea usando un escalar o entero usando un array.

A continuación, se muestra un ejemplo de su uso:

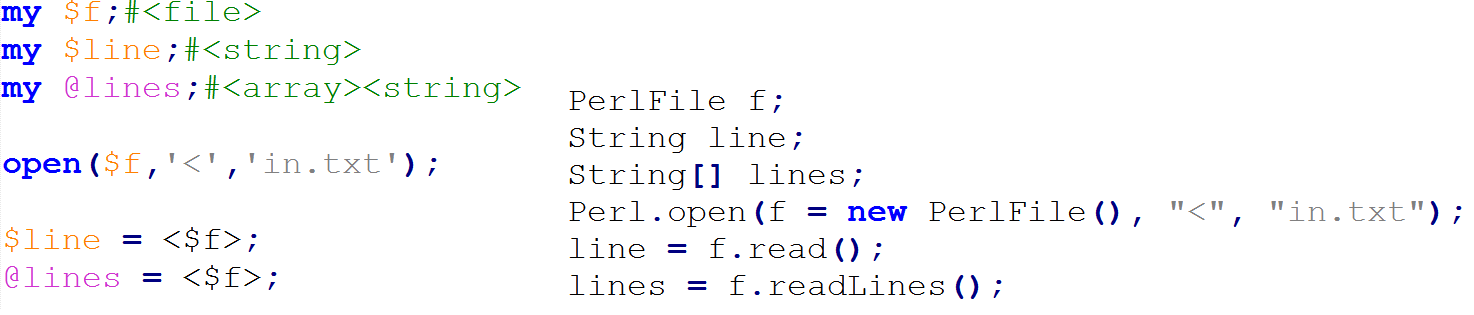


Figura 23: Ejemplo de lectura de fichero

Por otro lado, también es necesario especificar un formato para escribir en un fichero, a continuación, se muestra su uso usando la función print.

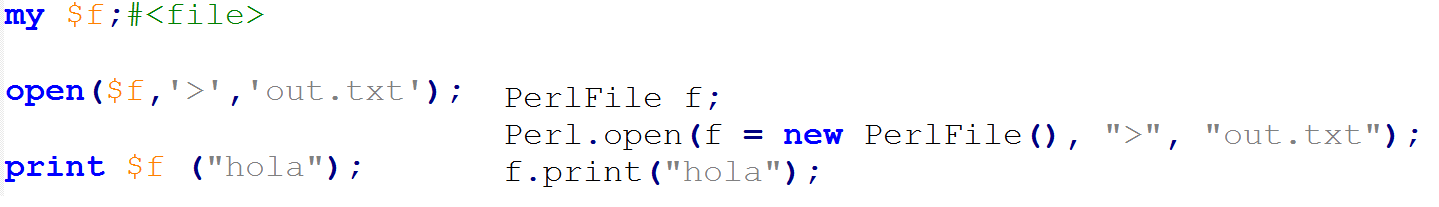


Figura 24: Ejemplo de escritura de fichero

Para terminar, es importante acordarse de cerrar los ficheros, para cerrar un fichero se llama a la función close, lo cual es tan trivial que no hace falta un ejemplo.

## Paquetes

El soporte para definir paquetes también esta soportado por el proceso de traducción. La inclusión de otros scrips se hace mediante la directiva ‘use’ que permite incluir un paquete definido en otro fichero.

El funcionamiento interno es muy simple, se requiere al programador que introduzca los ficheros de forma ordenada en el traductor. Luego según va finalizando el análisis de un fichero, el traductor conserva los todos métodos y variables globales en un hash usando como clave el nombre del paquete. De esta forma cuando se analiza el siguiente fichero, cuando se hace mención al paquete solo hace falta buscar en las las variables y métodos.

Un paquete corresponde a una clase java, las variables globales a atributos y los métodos a métodos estáticos. A continuación, se muestra un ejemplo práctico del funcionamiento de este proceso.

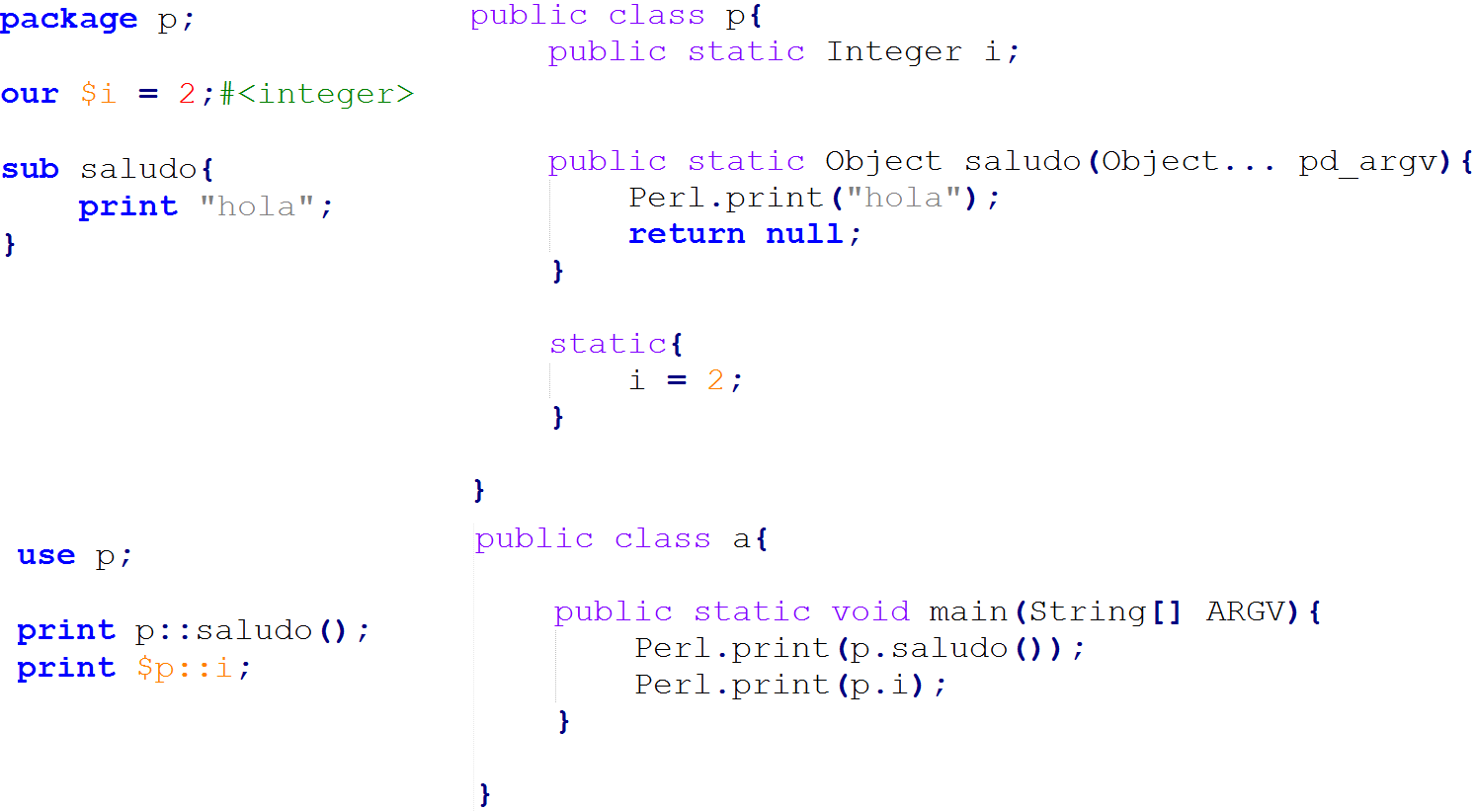


Figura 25: Creación y uso de un paquete con su traducción

## Funciones nativas

Para una mejor integración con Perl, es necesario dar soporte a las funciones nativas, para ello se han escogido las funciones más usadas y se ha definido una lista con ellas.

A continuación, se muestra una tabla con las funciones permitidas y una breve descripción del comportamiento de cada una de ellas.

|  |  |  |
| --- | --- | --- |
| Nº | Nombre | Descripción |
| 1 | chomp | Limpia una cadena y retorna el número de caracteres eliminados |
| 2 | chop | Quita el ultimo carácter a una cadena y lo retorna |
| 3 | close | Cierra un fichero y retorno su tuvo éxito |
| 4 | defined | Comprueba si una variable está asignada |
| 5 | delete | Borra un elemento de un Hash |
| 6 | die | Aborta el programa mostrando un mensaje de error |
| 7 | each | Recorre un Hashmap retornado de cada vez un clave-valor |
| 8 | exists | Comprueba si existe una clave en un HashMap |
| 9 | exit | Finaliza el programa |
| 10 | join | Une varias cadenas usando un separador |
| 11 | keys | Retorna las claves de un HashMap |
| 12 | lc | Transforma la cadena a minúsculas |
| 13 | lcfirst | Transforma la primera letra a minúsculas |
| 14 | length | Calcula la longitud de una candena |
| 15 | open | Abre un fichero |
| 16 | pop | Quita el último elemento de un array y lo retorna |
| 17 | print | Imprime |
| 18 | push | Añade un elemento al final de un array |
| 19 | shirft | Quita el primer elemento de un array y lo retorna |
| 20 | sort | Ordena un array |
| 21 | splice | Subdivide o remplaza segmentos de un array |
| 22 | split | Retorna una cadena dividida según un separador |
| 23 | substr | Subdivide o remplaza segmentos de una cadena |
| 24 | system | Realiza una llama al sistema |
| 25 | uc | Transforma la cadena a mayúsculas |
| 26 | ucfirst | Transforma la primera letra a mayúsculas |
| 27 | unshift | Añade un elemento al inicio de un array |
| 28 | values | Retorna los valores de un HashMap |

## Expresiones regulares

En este apartado se tratará el tema de la traducción de las expresiones regulares. En Perl las expresiones regulares, forman parte del código nativo sin necesidad de la invocación de ninguna función.

Las expresiones regulares pueden ser de tres tipos match, sustitución y transliteración y a su vez contener modificadores que pueden cambiar su comportamiento. Por ejemplo, existen modificadores para cambiar el ámbito de línea a multilinea o de caso no sensitivo acaso sensitivo entre otras.

Por esa razón se ha preferido conservar la máxima compatibilidad posible y dejar que sea Perl quien evalúe las expresiones regulares siendo invocado el intérprete desde java. Estas invocaciones, se llevan a cabo mediante funciones que invocan a la librería de Perldoop, por esa razón si en un futuro es posible ofrecer un soporte nativo java para la evaluación de expresiones regulares, no será necesario cambiar nada en el traductor.

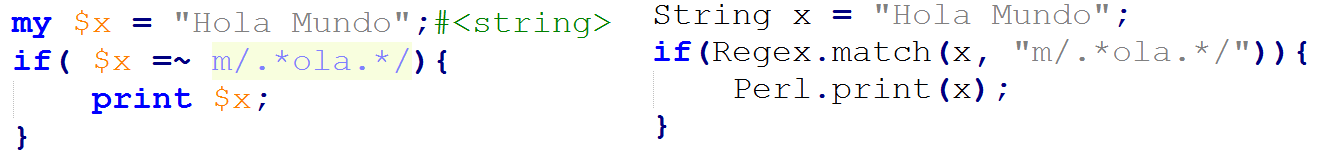


Figura 26: Expresión regular match

Como podemos ver en la figura 26, se envía a java la variable con la cadena a valida y la expresión regular escrita tal cual para poder reconstruirla en una llamada en el intérprete a Perl.

## Hadoop

En este apartado se explicará el desarrollo para poder convertir un código Perl secuencial en un código Java adaptado al paradigma MapReduce mejorando su ejecución en un clúster.

Para lograr este objetivo es necesario introducir nuevas etiquetas que permitan diferenciar un código Perl secuencial normal de uno que debe ser transformado. Por otro lado, también debe haber un compromiso por parte del programador, el manual define unas directrices en cuanto a la disposición del código para así poder aplicar la transformación. Por ejemplo, la versión secuencia de un Mapper, puede entenderse como un bucle que realiza el procesado secuencialmente.

## Mapper

El Mapper es el encargado de procesar un conjunto de datos, que por lo general se encuentran en forma de un fichero o un directorio y luego es almacenado en el sistema de ficheros Hadoop. El fichero es lo que recibe la función mapper fragmentado línea a línea, luego del procesado, los datos son convertidos en tupas que almacenan los datos en pares clave/valor.

En primer lugar, es necesario acotar el fragmento de código que pertenece al mapper separándolo del resto del código fuente. Para esto se usará la etiqueta con nombre <mapper\_code> seguida de un bloque que encerrará el código fuente. Con esta sintaxis, indicamos al analizador sintáctico que debe crear una clase que extienda de la clase Mapper, crear una sobrescritura de su método map y copiar la traducción en su interior.

En segundo lugar, la transformación de código secuencial a paralelo se realiza transformando un bucle que recibe sus datos de la entrada estándar. El criterio de transformación es eliminar el bucle por completo y asignar a la variable que obtenía los datos de la entrada estándar al argumento de la función map.

Por último, es necesario guardar los datos en algún lugar después del procesado, en Perl el criterio seguido es imprimir los parares clave valor usando dos separadores, uno para separar la clave del valor y otros para separar las tuplas. En Haddop, existe un elemento llamado contexto, en su interior existe una función llamada write que recibe como argumento esa tubla, es decir una clave y un valor. Para la transformación, emplearemos dos elementos, una llamada a la función print y la etiqueta <hadoop\_print>, la transformación se basa en el siguiente principio. Si obligamos al programador a imprimir sus datos usando el siguiente formato en la función print (clave, separador1, valor, separador2), es muy fácil realizar la transformación, debido a que solo debemos coger los argumentos 1 y 3 para luego pasárselos a la función write.

Para ilustrar todo los explicado en este apartado, se hara uso a continuación del famoso ejemplo WordCount en Perl junto con su traducción java.

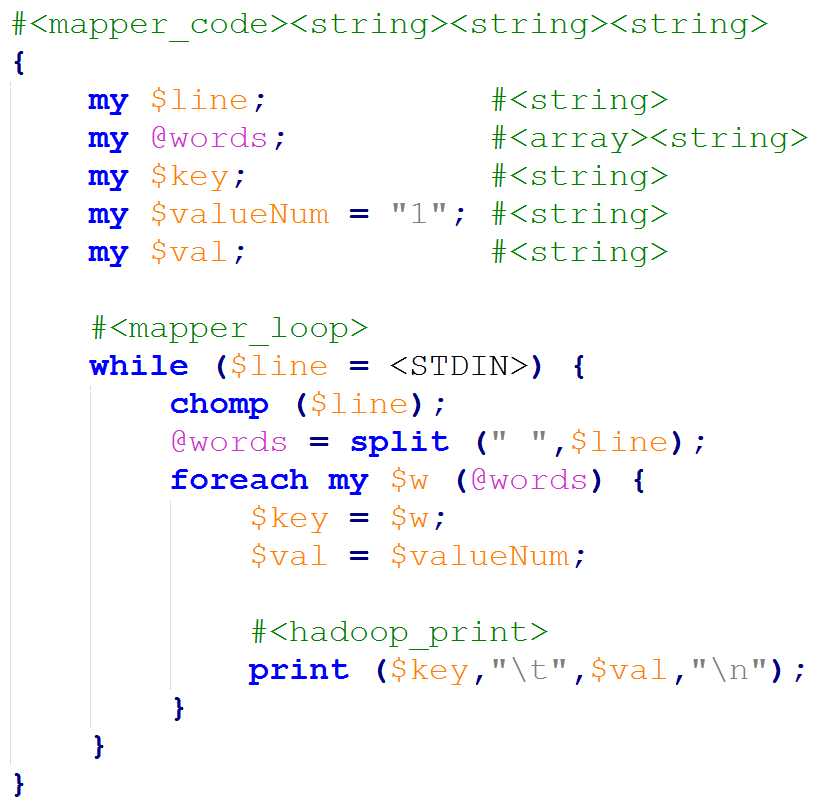


Figura 27: Mapper Perl

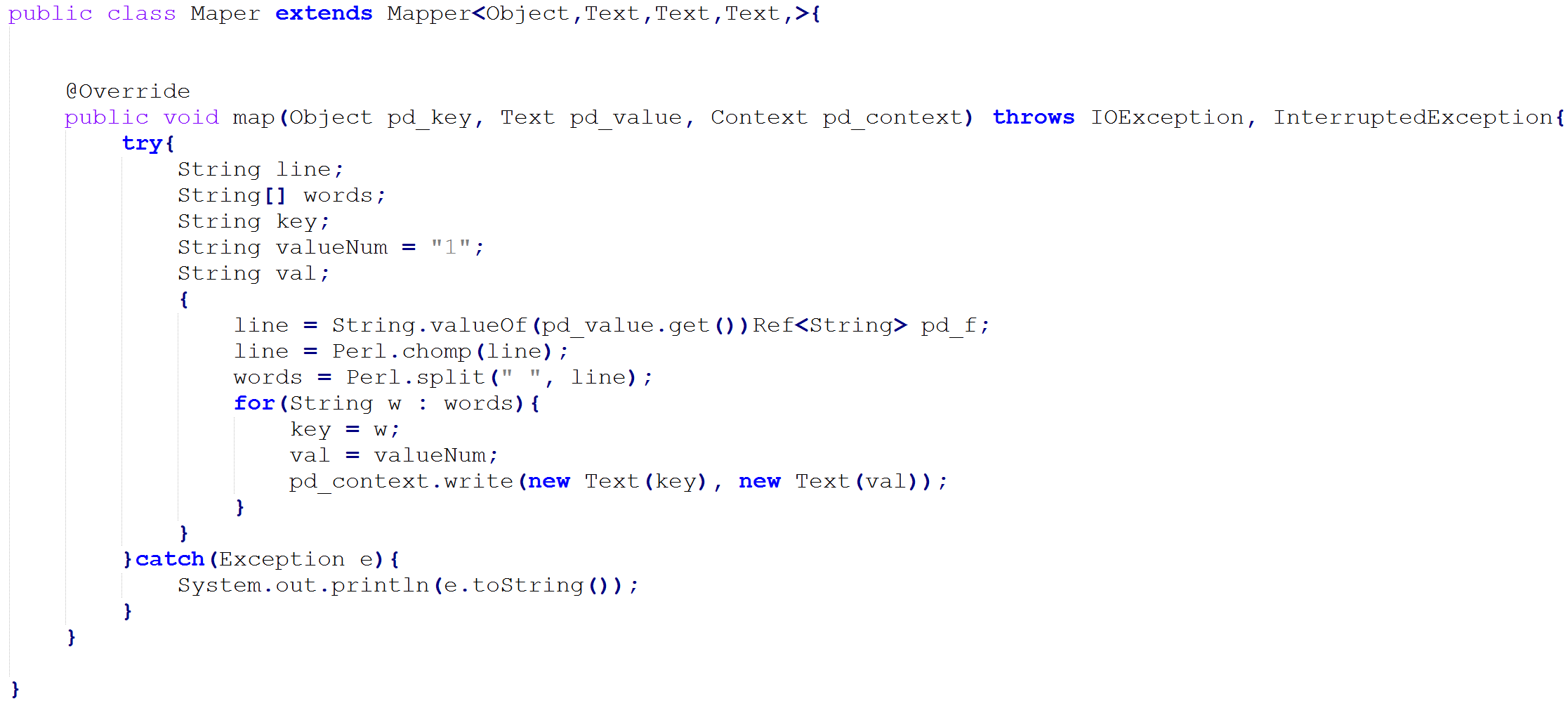


Figura 28: Mapper java

Como añadido final, junto a la etiqueta <mapper\_code>, hay definidos tres tipos. Esos tipos son opcionales y puede ignorarse, su función es cambiar el tipo de entra y salida del Mapper. Por defecto si no se especifica nada se establecerán todos los tipos como codenas(Text), el primer tipo hace referencia a cada línea leída del fichero, el segundo y el tercero hacen referencia a la clave y el valor respectivamente de la tupla de salida.

## Reducer

El Reducer es el encargado de combinar las tuplas clave/valor generadas por el Mapper con el objetivo de generar un resultado final. Una vez terminado se genera un nuevo conjunto de producción que es almacenado en el sistema de ficheros Hadoop.

Al igual que ocurrió con el Mapper, el Reducer también necesita etiquetas para adaptar el código fuente. El principal problema que presenta el Reducer frente al Mapper es que su sintaxis variable y es necesario proporcionar más libertad a la hora de su creación.

En primer lugar, debemos acotar el código del Reducer dentro de un bloque al igual que en el caso anterior, en este caso la etiqueta usada debe ser <reducer\_code> y el proceso de traducción es exactamente igual que el caso anterior cambiando el nombre de la clase padre y su método.

Para traducir los scripts de reducción, se han identificado tres regiones clave para transformar el código secuencial al Recucer hadoop. Una zona de definición de variables, la operación de reducción y por último las acciones una vez terminado con una clave y antes de empezar con la siguiente.

La primera región identificada con la etiqueta <reducer\_var> indica cuales son las variables que usaremos dentro del Reducer, la razón de esta definición es eliminar variables necesarias en Perl para aplicar transformaciones fuera de las operaciones de reducción y cambio de clave. En esta región es necesario identificar dos variables con una etiqueta especial cada una, <reducer\_key> y <reducer\_value> que corresponden a la variable que almacenara la calve y el valor respectivamente. Con esta región logramos identificar las variables a escribir en el código final y ademas las variables que habrá que inicializar con clave y valor para cada iteración sobre una tupla en cada reducción.

A continuación, con la etiqueta <reducer\_op> definimos la región que contiene las operaciones de reducción, o lo que quiere decir lo mismo, todas las acciones que son realizadas para todos los valores con la misma clave.

Para terminar, la etiqueta <reducer\_change> define la región con el código ejecutado al realizar todas las operaciones de reducción con una calve. En la mayoría de los casos esta zona es usada para guardar el resultado de la operación de reducción antes de empezar con la siguiente. Al igual que con el Mapper, el Reducer también puede hacer uso del contexto apra guardar sus datos, la sintaxis es idéntica a la del Mapper.

Al igual que el Mapper, el Reducer también puede cambiar el tipo de dato de su entrada y salida, que por defecto son cadenas. En este caso sería necesario definir cuatros tipos, uno más que el Mapper, los dos primeros para clave/valor d entrada y los otros dos para clave/valor salida.

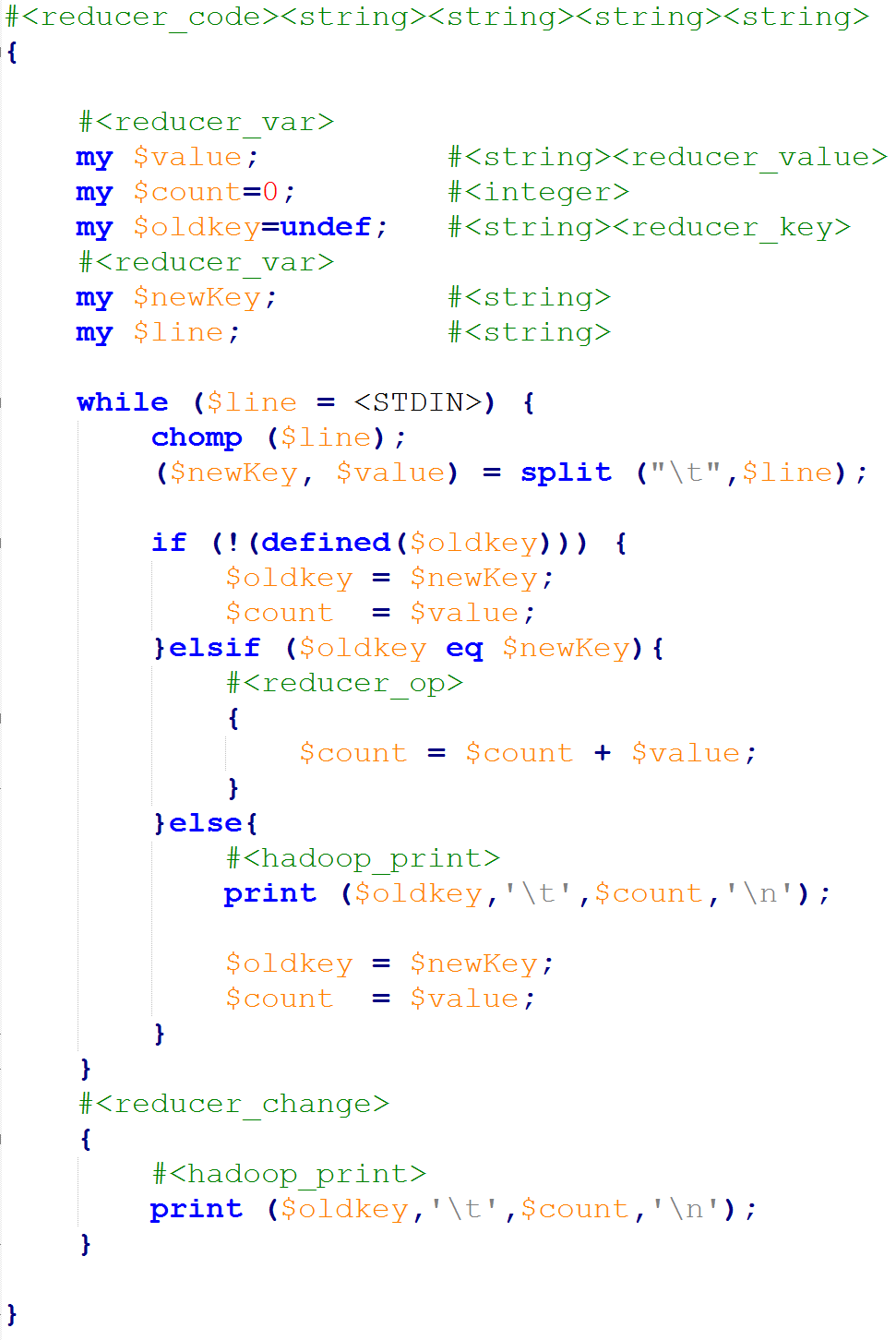


Figura 29:Reducer Perl

En la figura 29 podemos ver claramente la definición de un Reducer del famoso ejemplo WordCount que toma la salida de la operación Mapper y cuenta cuantos valores están asociados a cada clave.

Lo más importante antes de ver la traducción es darse cuenta de la importancia de la definición de regiones, si no fijamos la mayoría del código es necesario para limpiar líneas, separar clave de valor y gestionar el cambio de clave, lo cual en Hadoop ya no es necesario porque ya lo realiza el en nuestro lugar.

La traducción del Recuder es más compleja que la del Mapper, es este caso se almacena la traducción de cada uno de las regiones según se va procesando el fichero. Una vez terminado el bloque, se concatenan cada una de las regiones ignorando el código resultante de la traducción natural del bloque. Es importante realizar esta traducción porque, aunque el código secuencial ejecutado en paralelo produciría el mismo resultado, si nos fijamos ya en el código anterior, existen bloques condicionales que nunca se ejecutaran. Por ejemplo, oldkey eq newKey, si cada Reducer trabaja con una clave distinta, en el mismo Reducer nunca abra dos claves distintas y nuca se ejecutará el bloque correspondiente al else.

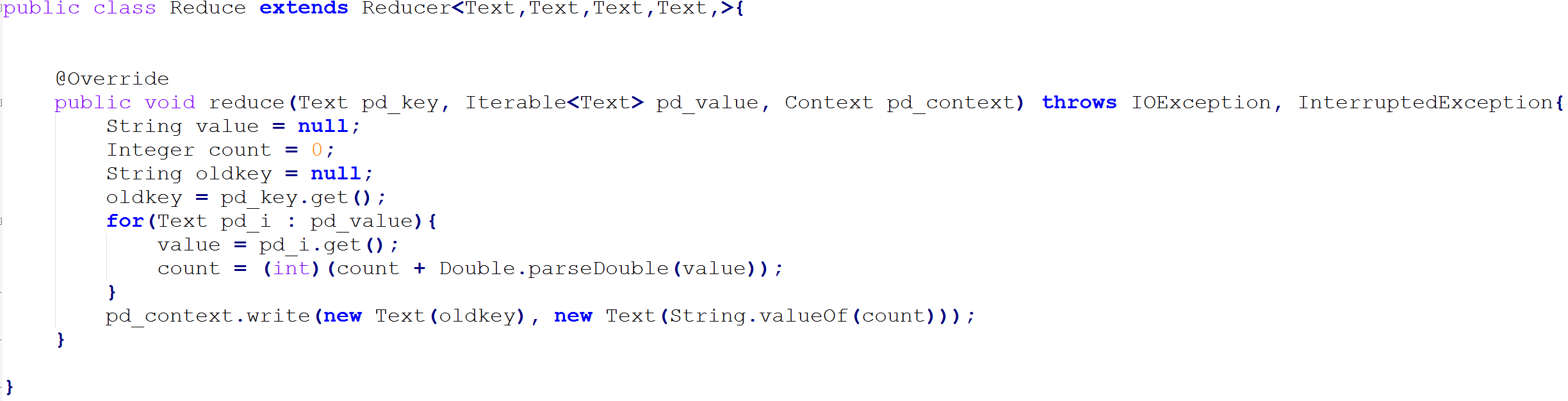


Figura 30: Reducer java

La figura 30 muestra un Reducer muy compacto comparado con su versión secuencial, si no fijamos es fácil apreciar donde se han situado cada una de las regiones definidas con etiquetas. Las variables han sido copiadas al inicio de la función sin aplicar ningún cambio significativo, la operación de reducción ha sido introducida dentro de un bloque que recorre todos los valores asociados a un calve y para terminar la operación de cambio de clave se realiza antes de finalizar la función Reducer.

## Librería java

Como se ha comentado en secciones anteriores, se ha creado una librería de soporte para el traductor, en este apartado se comentará su composición y las funcionalidades que contiene para hacer de soporte al traductor.

En primer lugar, la razón de la creación de una librería java es proporcionar un nivel de abstracción entre el comportamiento nativo de Perl y el código necesario para imitarlo en java. En otros casos era necesario la creación de una función para poder insertar una operación dentro de una expresión y en java son necesarias más operaciones como instanciación de objetos que obligarían a partir la expresión en varias líneas.

La librearía está formada por 9 clases que pueden ser agrupadas en dos tipos, el primer tipo ofrece un conjunto de funciones estáticas que son invocadas directamente de sin necesidad de ninguna instanciación. El segundo grupo son objetos que son instanciados para simular tipos de datos de Perl que no existen en java.

La implementación y diseño de la librería es muy simple y en esta sección se mostrará las función que realiza cada clase o los métodos que proporciona.

## Clases para Tipos

Las Clases para tipos imitan el comportamiento de tipos básicos para Perl que en java no existen como tal. La necesidad de estas clases es poder traducir una función sobre un tipo de dato por una función del mismo nombre y que realiza el mismo comportamiento. La otra razón importante es que la mayoría de las clases nativas Perl requieren generar una serie de Import para cada una de sus dependencias y de esta forma podemos ocultar todos ellos bajo la propia librería.

C:\Users\César\Desktop\Clases de Tipos.png

Figura 31: Clases para tipos

Como podemos ver en la figura 31, existen 5 clases para la gestión de los tipos de Perl: PerlList, HashPer y PerlFile. Las clases Fread y Fwrite solo existen para dar soporte a los ficheros actuando PerlFile como Fachada de ambos.

PerlFile y HashPerl existen solo para reproducir el comportamiento de Perl en la asignación. Cuando Perl guarda un valor en una lista hash, la asignación devuelve el mismo elemento para volver a usarlo, en java el comportamiento por defecto es devolver el elemento que existía en el mismo lugar antes de la asignación y en caso de no ser así devolver null. Las clases de la librería extienden las clases nativas de java heredando el método para insertar y retornando el valor que añaden dejando el comportamiento igual que en Perl.

Por otro lado, PerlFile, es el tipo de dato creado cuando se define una variable como fichero, luego según el uso que se le de a ese fichero, se creara un Fread si se abre para lectura o Fwrite si es para escritura. Los métodos de cada uno de ellos siguen la filosofía de trabajo con fichero en java, primero abren el fichero y luego crea un buffer para aumentar el rendimiento, el proceso es realmente sencillo y no requiere de mas acciones.

## Clases para Funciones

Las clases para funciones son un conjunto de métodos estáticos agrupados en tres clases Regex, Pd, Perl.

La clase Regex contiene la funcionalidad para la ejecución de las expresiones regulares, es decir, un método para match, otro para sustitución y otro para transliteración. En cada función, se realiza una llamada al sistema para que Perl las evalué la expresión y el resultado es retornado por la función. Este sistema fue adoptado por el tiempo requerido para estudiar toda la sintaxis de las expresiones regulares de Perl y buscar una librería Java para imitarlas. En el futuro, si se encuentra una solución mejor, solo habría que remplazar el código dentro de las funciones sin más complicaciones.

La clase Pd, contiene las funciones relativas a comportamientos de Perl que no existen en java de forma nativa o que necesaria evaluar la expresión en tiempo de ejecución debido a que no se puede llevar a cabo en tiempo de ejecución.

|  |  |
| --- | --- |
| Nombre | Descripción |
| toInteger | Castea un número a integer\* |
| toLong | Castea un número a Long\* |
| toFloat | Castea un número a Float\* |
| toDouble | Castea un número a Double\* |
| repeat | Operador Perl x que repite una cadena un numero de determinado de veces. |
| and | Realiza lo operación and con números enteros, retornando el segundo si el primero es falso o el primero en caso contrario. |
| or | Realiza lo operación and con números enteros, retornando el primero se es cierto o el segundo si el primero es falso. |
| xor | Realiza la opción lógica xor.\*\* |
| hash | Crea un hash desde un conjunto de expresiones. |
| range | Operador Perl .. que crea una lista entre dos intervalos que pueden ser números o cadenas. |
| copy | Realiza una copia superficial de una colección, usado cuando se igualan dos colecciones. |
| array | Crea un array desde una lista. |
| eval | Comprueba si una expresión es cierta. \*\* |

Tabla 8: Tabla de funciones auxiliares librería

\*Las funciones de casting son necesarias ante la imposibilidad de convertir un numero sin saber si es un tipo primitivo o un objeto. Las cuatro funciones reciben como argumento un Number que representa cualquier número y al cual java puede convertir cualquier tipo número sea del tipo que sea, luego cada función retorna intValue(), longValue(), floatValue() o doubleValue().

\*\*Tanto eval como xor no pueden escribirse en el código porque requieren duplicar la expresión para realizar la operación y en algunos casos no podemos asegurar que la misma no implicaría un alto coste o que se obtendrá el mismo resultado. La función xor tiene que realizarse mediante or y and y en el caso de validar si algo es cierto es necesario realizar múltiples comprobaciones, por ejemplo, para la cadena se necesita comprobar si no es nula, vacía o es “0”. En el caso de operaciones del estilo “x++” o retornos de funciones, nunca se podrá replicar la expresión.

Por ultimo queda la clase Perl, esta clase contiene todas las funciones nativas especificadas en el apartado 8.4.7. Su implementación no requiere diseño alguno, toda pueden realizarse mediante objetos y funciones ya definidas en Perl y que por claridad no deberían escribirse directamente en la traducción. Por ejemplo, si un script de Perl contiene ‘print(“Hola mundo”)’ la traducción seria ‘Perl.print(“Hola mundo”)’, si un programador de Perl mira el código java, le será más fácil de entender que ver ‘System.out.print(“Hola mundo”)’.

# Pruebas

# Bibliografía

1. Perldoop [En línea]. Available: https://github.com/citiususc/perldoop
2. Tom Christiansen, brian d foy & Larry Wall with Jon Orwant, Programming Perl 4h Edition, O’REILLY
3. Tom White, Hadoop The Definitive Guide, O’REILLY
4. Manual Perldoop, Sección 2.2 [En línea]. Available: https://github.com/citiususc/perldoop/blob/master/doc/UserManual\_Perldoop\_EN.pdf
5. Manual Perldoop, Sección 2.3 [En línea]. Available: https://github.com/citiususc/perldoop/blob/master/doc/UserManual\_Perldoop\_EN.pdf
6. PMBOK Guides,A Guide To The Project Management Body Of Knowledge Project, Management Institute, 2004.
7. IEEE830. recommended practice for software requirements speciﬁcations. Technical report, 1998
8. Douglas Bell. Software Engineering for Students. Addison-Wesley, 2005.
9. Atlassian Software, «BitBucket,» [En línea]. Available: https://bitbucket.org/
10. PLY (Python Lex-Yacc) [En línea]. Available: http://www.dabeaz.com/ply/
11. InfoJobs, «InfoJobs,» [En línea]. Available: https://www.infojobs.net/.
12. Eclipse IDE [En línea]. Available: https://eclipse.org/
13. Netbeans IDE [En línea]. Available: https://netbeans.org/
14. SourceTree [En línea]. Available: https://www.sourcetreeapp.com
15. Draw.io [En línea]. Available: https://www.draw.io/
16. Perl 5 documentation [En línea]. Available: http://perldoc.perl.org/perlop.html