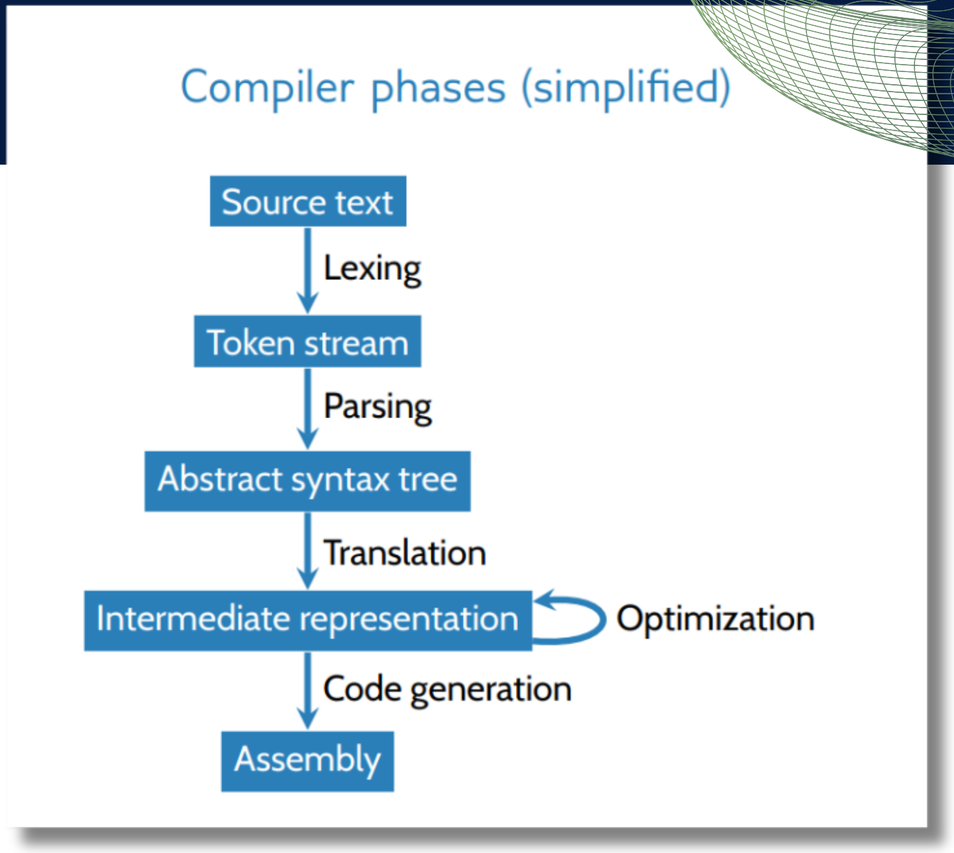
¿Qué es un lexer?

Se trata de la primera fase en todos los compiladores modernos donde se escanea y lee código fuente de tal manera que pueda convertir sus entradas en lo que se conoce como **Token Stream** que da paso a la creación de un **Árbol sintáctico** que guía el resto de la compilación.



Ejemplo

En la primera imagen podemos observar un código fuente bastante básico en GO, por otro lado, la siguiente imagen muestra la respuesta de un lexer. Como vemos, cada valor corresponde a un token que es reconocido por el lexer.

```go
package main

import "fmt"

func main() {
    fmt.Println("Hello world!")
}
```

Imagen 1.

```
1:1     package "package"
1:9     IDENT   "main"
1:13    ;       "\n"
3:1     import  "import"
3:8     STRING  "\"fmt\""
3:13    ;       "\n"
5:1     func    "func"
5:6     IDENT   "main"
5:10    (       ""
5:11    )       ""
5:13    {       ""
6:2     IDENT   "fmt"
6:5     .       ""
6:6     IDENT   "Println"
6:13    (       ""
6:14    STRING  "\"Hello world!\""
6:28    )       ""
6:29    ;       "\n"
7:1     }       ""
7:2     ;       "\n"
```

Imagen 2.

Definiciones Básicas

Token

Se trata de una agrupación de caracteres que constituyen los símbolos que forman las sentencias del lenguaje tales como: palabras reservadas, identificadores, operadores, símbolos especiales.

Lexema

Secuencia de caracteres que coinciden con determinado token como el nombre de un identificador o el valor de un numero.

Patrón

Es la forma en que se describen los tipos de lexema usando expresiones regulares.

Token (Componente léxico)	Lexema	Patrón
While	While	While
Suma	+	+
Identificador	a, valor, b	[a-zA-Z]+
Número	5, 3, 25, 56	[0-9]+(\.[0-9]+)?

VENTAJAS

1ª Ventaja

Se simplifica el diseño, puesto que hay una herramienta especializada en el tratamiento del fichero que contiene el código fuente.

2ª Ventaja

Aumenta la portabilidad del compilador, pudiendo tenerse versiones diferentes para distintos formatos del texto de código fuente (ASCII, EBCDIC, etc.).

3ª Ventaja

Mejora la eficiencia al ser una herramienta especializada en el tratamiento de caracteres.

4ª Ventaja

Detección de determinados errores fáciles de corregir a este nivel (5.25 por 5,25).

ERRORES LÉXICOS

Los errores léxicos son detectados, cuando durante el proceso de reconocimiento de caracteres, los símbolos que tenemos en la entrada no concuerdan con ningún patrón. Hay que tener en cuenta que hay pocos errores detectables por el analizador léxico, entre ellos están:

Nombres incorrectos de los identificadores

Se debe a que se utilizan caracteres inválidos para ese patrón, como por ejemplo un paréntesis, o se empieza por un número.

Números incorrectos

Debido a que está escrito con caracteres inválidos (puntos en lugar de comas) o no está escrito correctamente.

Palabras reservadas escritas incorrectamente

Se producen errores de ortografía. El problema aquí es cómo distingues entre un identificador y una variable reservada.

Caracteres que no pertenecen al alfabeto del lenguaje

Ejemplos: @, €, ¿, ?, ñ, etc.