Guión de la exposición

- Motivación
- Estado del arte
- Compresores semiestáticos
- Compresores dinámicos
- Compresores dinámicos Ligeros
- Compresión variable-to-variable

• Compresión de colecciones crecientes: PB-ETDC
• Mejora ratio compresión:
  • P-ETDC
  • RP-ETDC

Variable-to-variable: Mejora ratio compresión

Motivación

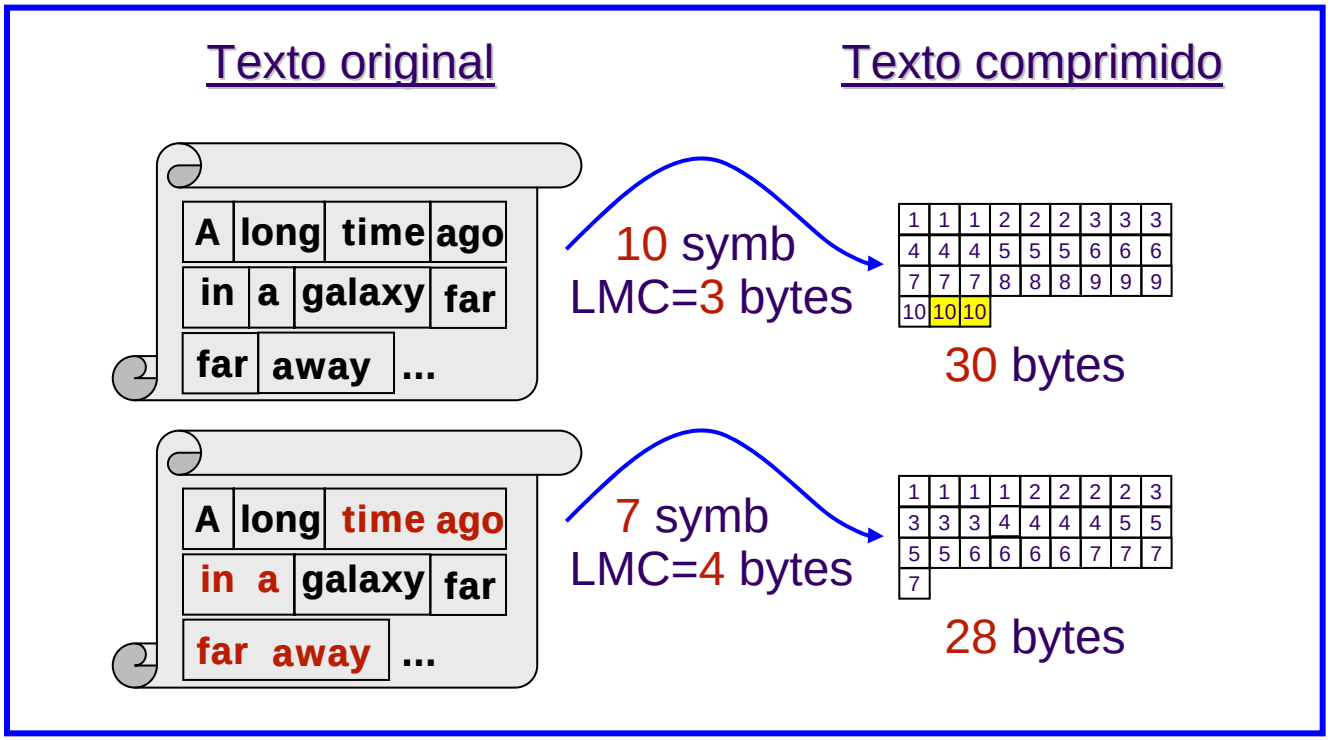
- **Mejorar el ratio de compresión obtenido por técnicas semi-estáticas como PH, SCDC, ETDC (≥ 31%).**

- **Mantener la posibilidad de**
  - Buscar eficientemente texto comprimido
  - Descompresión aleatoria

- **➔ Permitamos que el vocabulario de ETDC (semistático) admita pares de palabras y no sólo palabras simples. → PETDC**

Variable-to-variable: Mejora ratio compresión

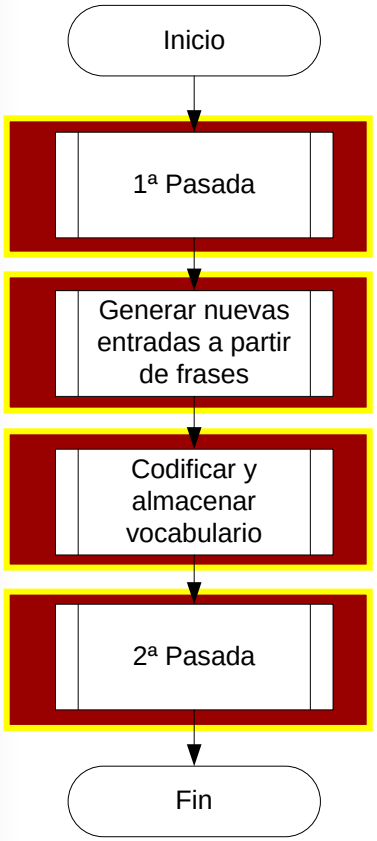
P-ETDC



- Tamaño final = numCódigos * LMCodigos
  - Idea ➔ reduzcamos el número de códigos generados!!

Variable-to-variable: Mejora ratio compresión

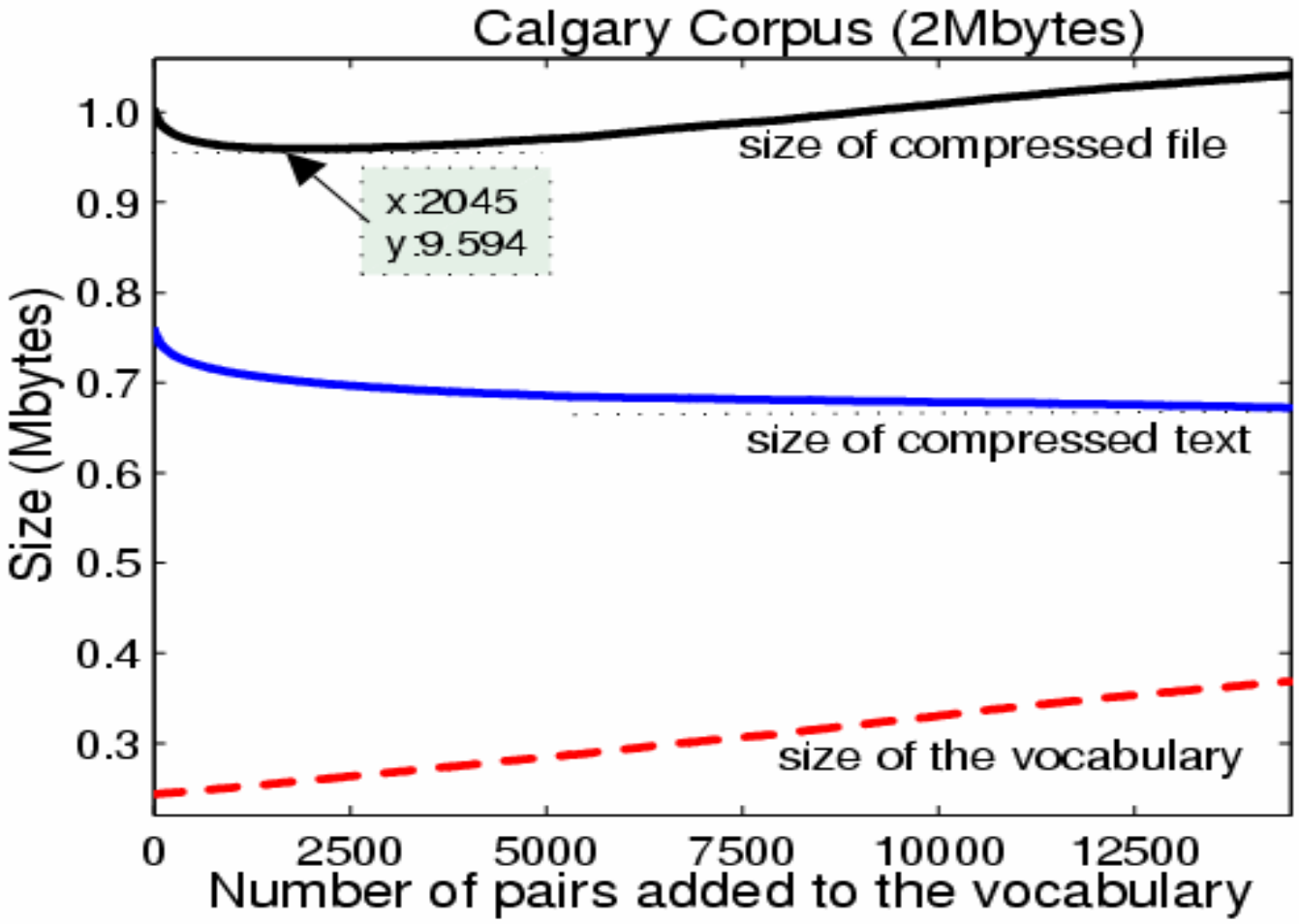
P-ETDC



- **Funcionamiento del compresor:**
  - **1ª pasada**
    - **Obtener palabras y pares de palabras posibles**.
    - Ordenar pares por frecuencia
  - Recorrer pares generados
    - Evaluar si mejoran compresión
      - Heurística
      - Añadir par **ab** → **descartar** pares x**a**, **b**y
    - Evaluar todos pares
  - **Codificación** vocabulario usando ETDC.
  - **2ª pasada**
    - Sustitución símbolo (par o palabra) por código
    - Almacenar vocabulario.

Variable-to-variable: Mejora ratio compresión

PETDC: Selección de pares



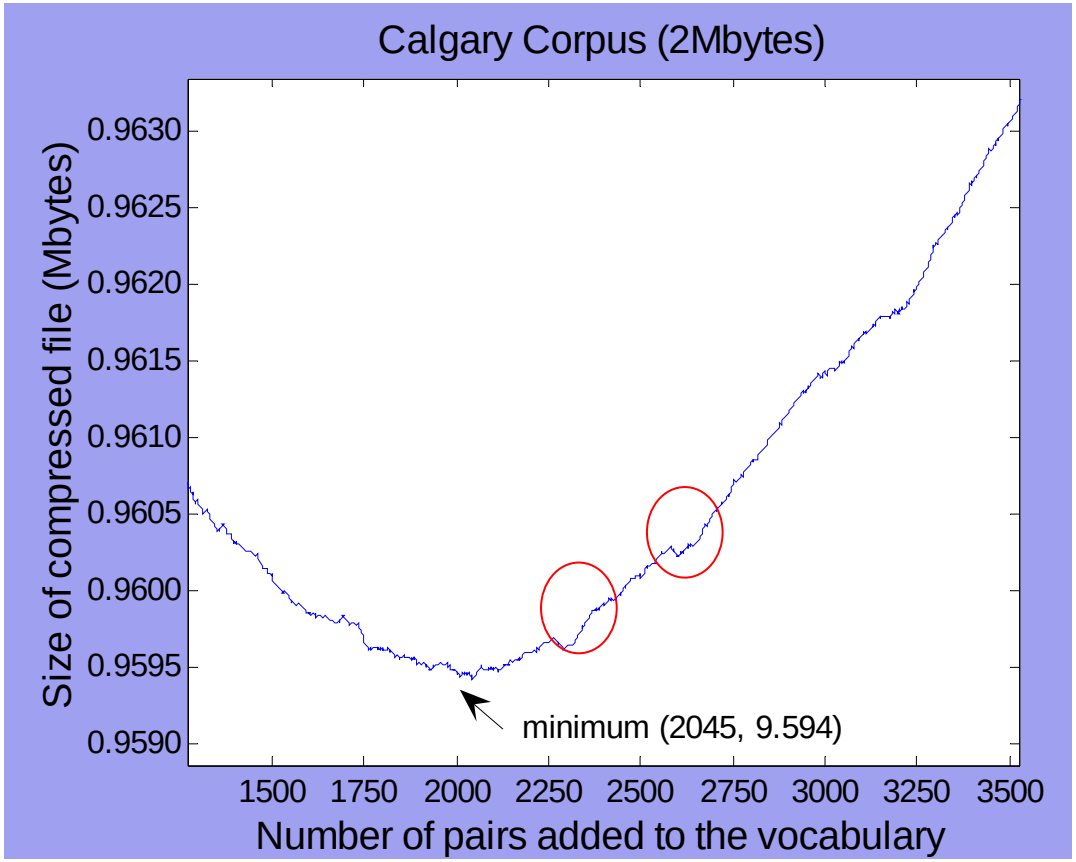
Variable-to-variable: Mejora ratio compresión

PETDC: ¿función parada? ➔ NO

Función de parada ??

¿parar cuándo no mejoremos?

Pero hay mínimos locales…



Evaluar todos pares ➔ **no usar función de parada**

Variable-to-variable: Mejora ratio compresión

PETDC: Selección de pares

- Heurística para decidir si un par $\alpha\beta$ debe ser añadido …
  - Tamaño texto comprimido antes de añadir par $\alpha\beta$
  - Tamaño texto comprimido después de añadir par $\alpha\beta$
  - Incremento tamaño vocabulario para incluir $\alpha\beta$

$$skip_{bytes} = f_\alpha * |C_\alpha| + f_\beta * |C_\beta|$$

$$add_{bytes} = f_{\alpha\beta} * |C_{\alpha\beta}| + (f_\alpha - f_{\alpha\beta}) * |C'_\alpha| + (f_\beta - f_{\alpha\beta}) * |C'_\beta| + K$$

Variable-to-variable: Mejora ratio compresión

PETDC: Elección de pares

Vocabulario

	Frec	Palabra
1	100	doctor
2	90	house
	…	…
128	85	
129	81	
	…	
	70	sunny
200		

Código asignado

10000000	
10000001	
…	
11111111	
00000000	10000000
…	….
00000000	10001010

1 BYTE

2 BYTES

house + sunny = 90 + 70 * 2 = 230 bytes

Variable-to-variable: Mejora ratio compresión

PETDC: Elección de pares

Vocabulario

Frec Palabra

Código asignado

Sin par “sunny house”

230 bytes

Con par “sunny house”

200 bytes

Ganancia de compresión de 30 bytes

sunny + house = 130	cluster (sunny house) = (60 + 30 = 160) bytes	200 bytes

Variable-to-variable: Mejora ratio compresión

PETDC: Elección de pares

Vocabulario

Frec word

Código asignado

Sin par “doctor house”

160 bytes

Con par “doctor house”

220 bytes

Pérdida de compresión de 60 bytes

doctor + house = 100 + 30 * 2 = 160 bytes

doctor + house + (doctor house) = (80 + 20 + 10) * 2 = 220 bytes

Variable-to-variable: Mejora ratio compresión

PETDC: Ejemplo

A D C B A C D C C D A B A B A C D C

Vocabulario

Texto	frec.
A	7
D	5
C	6
B	3
BA	3

Pares posibles

BA	3
DB	1
CD	1
AD	2
DA	2
AC	2
AB	2
CC	1
CB	1
CA	1

Primera pasada: Contabilizar frecuencia palabras y detectar pares

Ordenar por el vocabulario

Crear nuevo vocabulario

Variable-to-variable: Mejora ratio compresión

PETDC: Ejemplo

A D C B A C D C C D A B A B A C D C

Vocabulario

Texto	fre.
A	4
D	3
C	6
B	0
BA	3
DC	3

Pares

BA	3
DC	3
CD	3
AD	2
DA	2
AC	2
AB	2
CC	1
CB	1
CA	1

Continuar hasta evaluar todos los pares posibles

Descartar pares de palabras de baja frecuencia

Variable-to-variable: Mejora ratio compresión

PETDC: Ejemplo

A D C B A C D C C D A B A B A C D C

Vocabulario

Texto	fre.	**Código**	
A	4	100	
DC	3	101	
C	3	110	
BA	3	111	
D	2	000	100
B	0		

Cabecera

A\0 **000100 110** C\0 **000101 110** D\0 B\0

Ordenar vocabulario y Segunda pasada generar codificación

Texto codificado

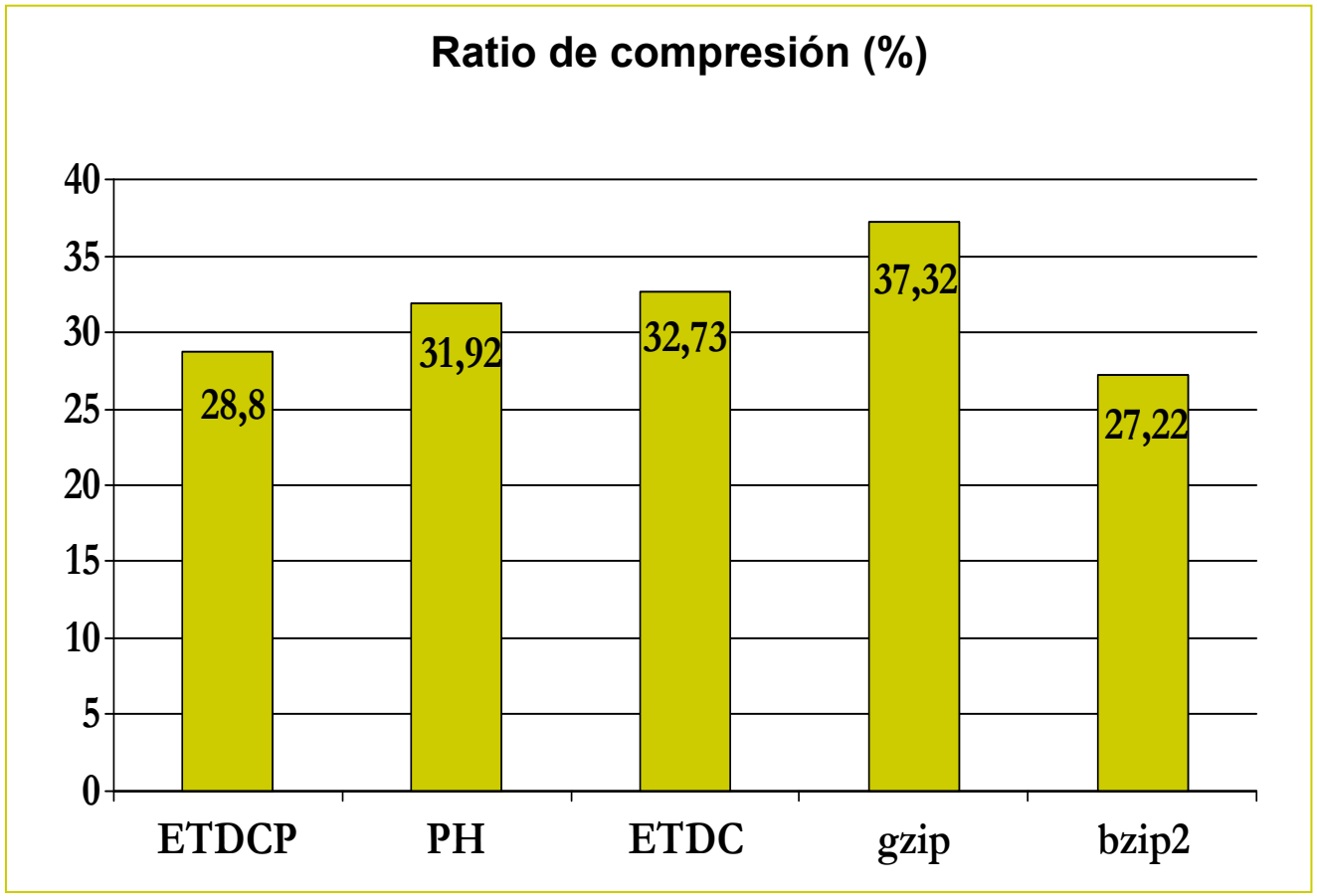
100101111110000100110110000100110111111100101

Fichero comprimido

Variable-to-variable: Mejora ratio compresión

P-ETDC: Resultados Empíricos

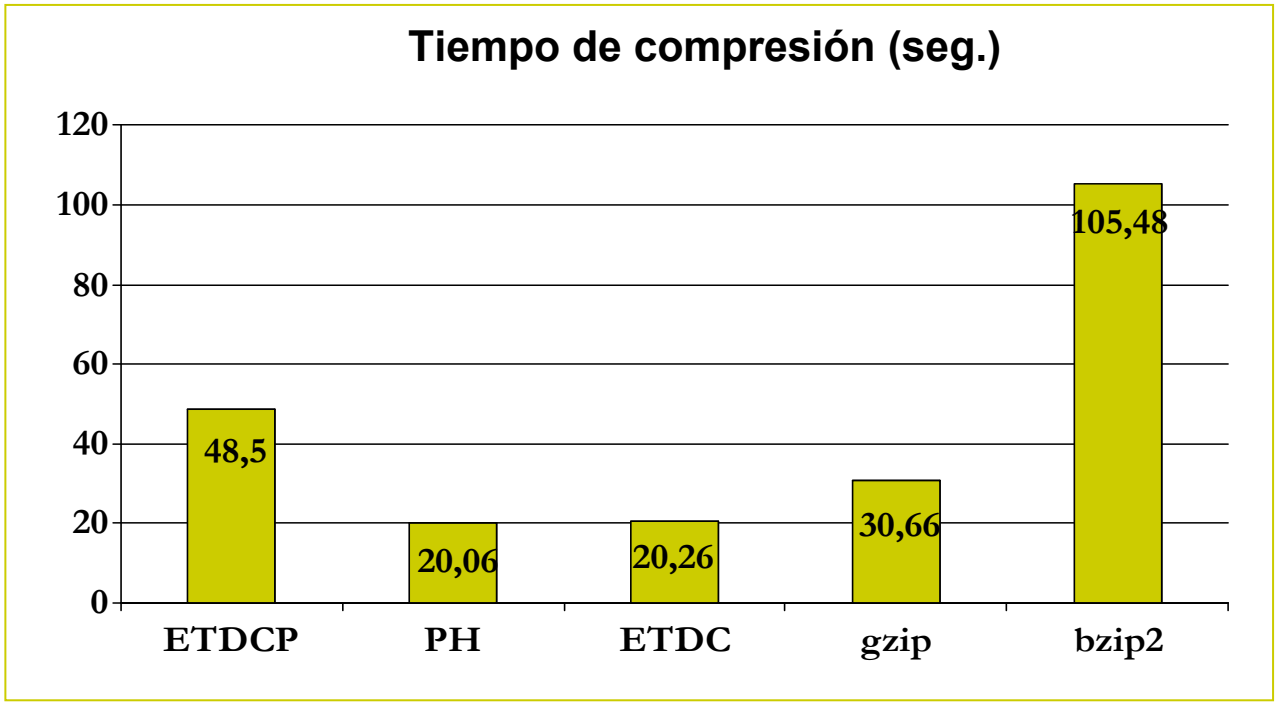
Corpus AP



Variable-to-variable: Mejora ratio compresión

P-ETDC: Resultados Empíricos

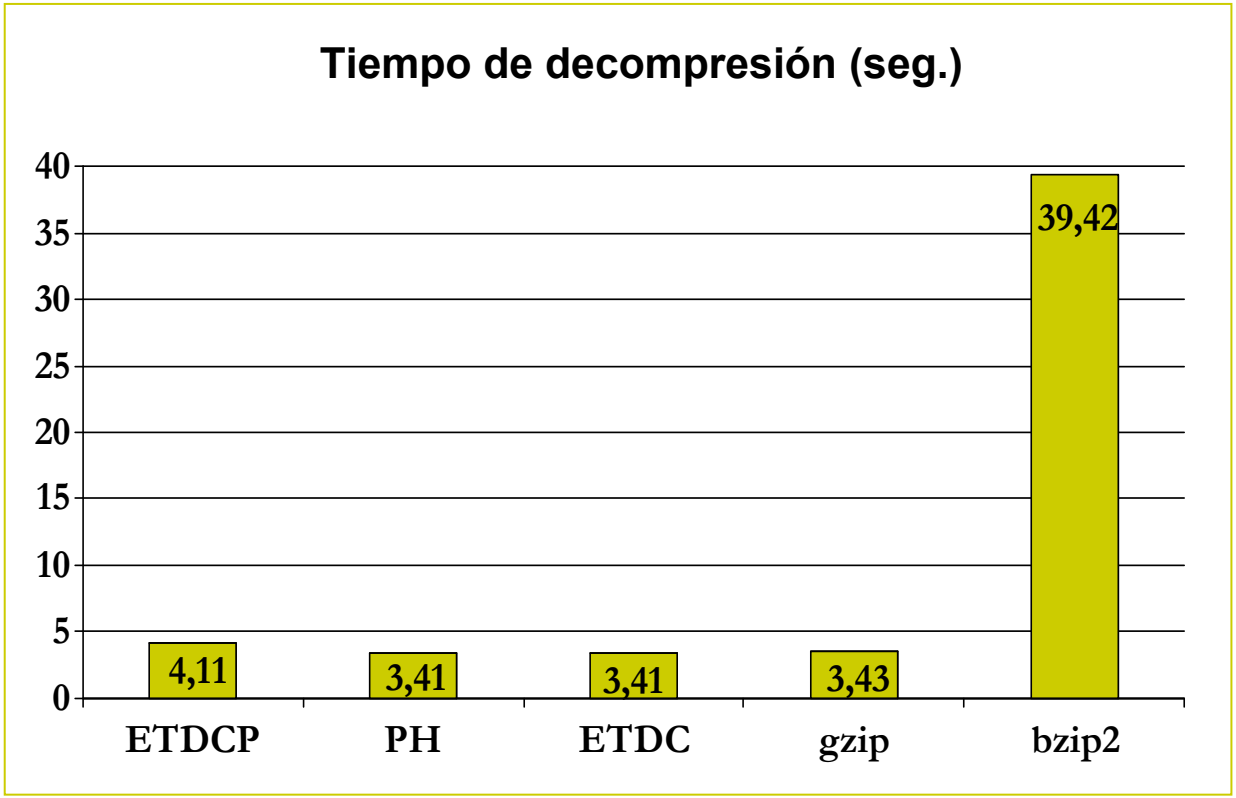
Corpus AP



Variable-to-variable: Mejora ratio compresión

P-ETDC: Resultados Empíricos

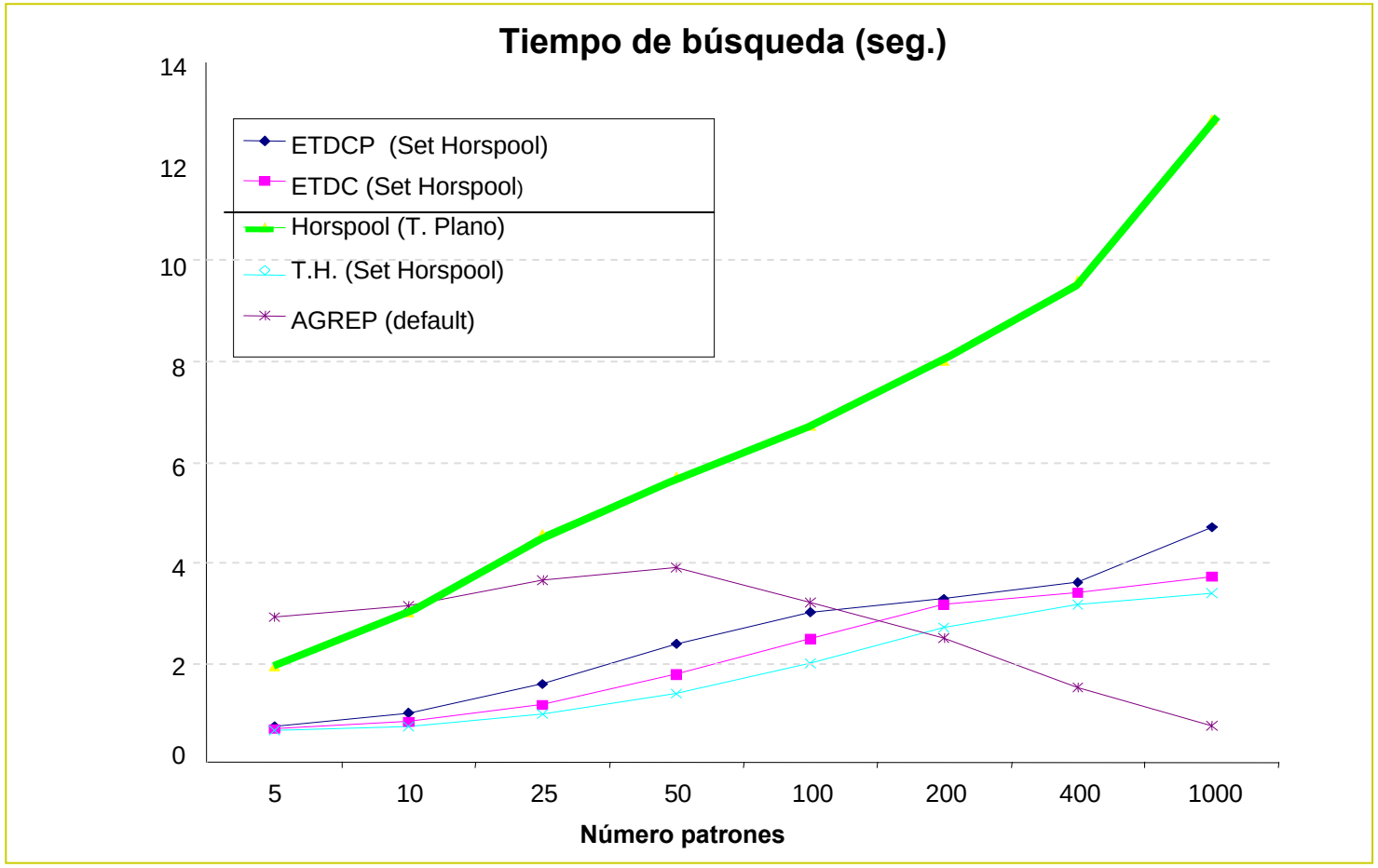
Corpus AP



Variable-to-variable: Mejora ratio compresión

P-ETDC: Resultados Empíricos

Corpus AP



Guión de la exposición

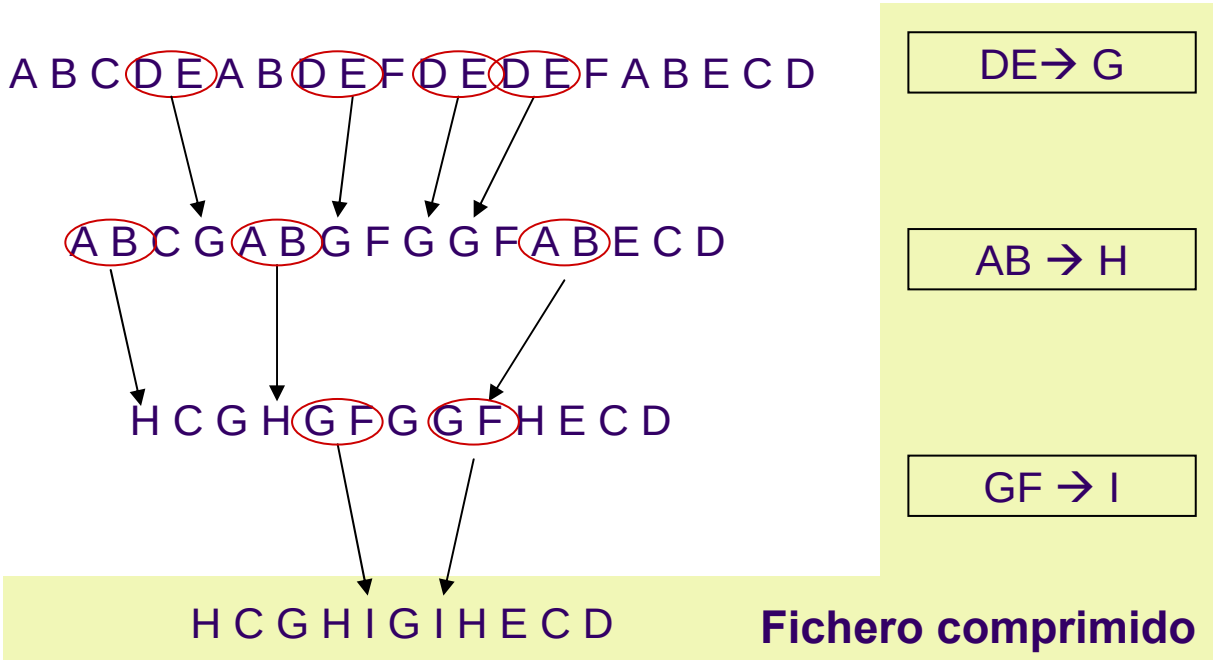
- Motivación
- Estado del arte
- Compresores semiestáticos
- Compresores dinámicos
- Compresores dinámicos Ligeros
- Compresión variable-to-variable

• Compresión de colecciones crecientes: PB-ETDC
• Mejora ratio compresión:
  • P-ETDC
  • RP-ETDC

Variable-to-variable: Mejora ratio compresión

Recursive PETDC (RPETDC)

- **Idea: Generalizar uso de PETDC**.
- Basado en Byte Pair Encoding (BPE), (orientado palabras Wan'2003)
  - Sustituir pares de símbolos por uno nuevo (pares, tríos,…)



- **RPETDC:**
  - **1º:** Generar pares recursivamente (heurísticas)
  - **2º**: comprimir con ETDC (la secuencia comprimida) … u otra técnica estadística
  - **Finalmente**: compactar el vocabulario

Variable-to-variable: Mejora ratio compresión

RPETDC: Resultados empíricos

Corpus CR