

1. Introducción

PETRAtag es un etiquetador morfosintáctico que ha sido desarrollado por el departamento de investigación de Equus Traducciones con la ayuda de la Universidad de Granada (Grupo de investigación PETRA HUM 815). Esta aplicación asigna a cada palabra de un texto un lema y una etiqueta que indica sus rasgos morfosintácticos, es decir, nos permite conocer para cada palabra de un texto, si se trata de un nombre o un verbo o cualquiera que sea su categoría morfosintáctica, así como su género, número, etc.

Este proceso permite realizar operaciones avanzadas con textos, tales como detectar los regímenes preposicionales de verbos y realizar muchas otras búsquedas complejas que posibilitan la creación de traducciones que se ajustan con total precisión a los requisitos solicitados.

2. Instalación

PETRAtag se ha creado utilizando Visual Basic .NET. Por tanto, para ejecutar esta aplicación es imprescindible haber instalado en el equipo el Microsoft .NET Framework. Se trata de un programa completamente gratuito y redistribuible que puedes encontrar en la misma página de donde has descargado este manual. Sólo tienes que descargarte el archivo llamado dotnetfx.exe y aparecerá un asistente que te guiará por los pasos necesarios para instalarlo.

Una vez instalado el Microsoft .NET Framework ya puedes instalar y ejecutar correctamente el programa propiamente dicho. Para ello, descarga el archivo PETRAtag.zip y descomprímelo en un directorio temporal. A continuación, ejecuta el programa Instalar.exe, que iniciará el asistente que te ayudará a instalar el programa en la ubicación que desees. Una vez instalado, deberá haber aparecido en el escritorio un icono PETRAtag con el que podrás ejecutar esta aplicación.

3. Primeros pasos

Para comenzar vamos a cargar un texto y etiquetarlo, para ver los resultados que obtenemos. Para ello, elegimos en el menú **Archivo**, la opción **Abrir** que hará que aparezca un cuadro de diálogo en el que podemos elegir el archivo que deseamos etiquetar. Como ejemplo, vamos a abrir el archivo **Becquer, Gustavo Adolfo - El monte de las ánimas.txt**, que se incluye junto con la instalación de este programa. Una vez que seleccionemos el archivo y hagamos clic en **Abrir**, PETRAtag abrirá automáticamente el archivo y lo etiquetará. Según la longitud del texto, este etiquetado puede ser casi instantáneo o puede tardar varios minutos. Como indicación de este proceso, en la barra de estado que hay en la parte inferior de la pantalla principal, aparece un pequeño mensaje que nos indica las líneas que ya se han etiquetado. Al finalizar este proceso, aparecerán unos resultados similares a los siguientes:

Hay un total de 6351 elementos.
Se han etiquetado 6006 palabras (94.5677%).

Hay 0 etiquetas.
Hay 0 números.
Hay 928 signos de puntuación.
Hay 128 nombres propios.

La primera línea de estos resultados indica el número total de elementos o tokens que hay en el texto. Hay que tener en cuenta que esta cifra incluye no sólo a las palabras sino también a las cifras y los signos de puntuación. Además, hay determinados grupos de palabras que, en las ocasiones en las que funcionan como una única unidad, se etiquetan juntas, como ocurre con la locución "en vano", que frecuentemente funciona como una locución adverbial.

La siguiente línea nos indica el número de estas palabras que se han logrado etiquetar. El diccionario actual de PETRAtag tiene cerca de 9000 raíces, lo que le permite etiquetar prácticamente en todas las ocasiones más del 90% de cualquier texto y, normalmente, más del 95%. Este diccionario se sigue actualizando regularmente, por lo que puedes consultar la página de donde descargaste la aplicación para obtener una versión más reciente y completa del diccionario. Otra posibilidad consiste en introducir personalmente las palabras que desees etiquetar como verás en el apartado **Introducción de nuevas palabras en el diccionario**.

Las siguientes líneas nos indica el número de etiquetas (cadenas de texto comprendidas entre los caracteres "<" y ">"), cifras, signos de puntuación y nombres propios que incluye el texto.

Ahora que ya tenemos etiquetado el texto, podemos pulsar Ctrl+L para ver cómo ha quedado etiquetado el texto. También podemos hacer esto, seleccionando en el menú **Ver** la opción **Ver archivo etiquetado y Lista**. A continuación, vemos un pequeño ejemplo del etiquetado que debemos haber obtenido.

```
La el da0fs0
Noche noche ncfs000
de de sps00
Difuntos difunto aq0mp0
, , Fc
me me pp1cs000
despertó despertar vmis3s0
a a sps00
no no rn
sé saber vmip1s0
qué qué dt0cn0
hora hora ncfs000
el el da0ms0
doble doble aq0cs0
de de sps00
las el da0fp0
campanas campana ncfp000
.. Fp
```

Como vemos, junto a cada palabra aparece su lema (la forma canónica) y una etiqueta que indica sus características morfosintácticas. Para ver el significado de estas etiquetas podemos consultar el apéndice A, que explica el significado exacto de cada etiqueta. Este sistema de etiquetas ha sido desarrollado por por el grupo CLiC -Centre de Llenguatge i Computació- de la Universitat de Barcelona.

Ahora, para comenzar a aprovechar las posibilidades del etiquetado, vamos a extraer todos los adjetivos que hay en este texto con vistas a un posible estudio estilístico del

texto con textos de corpus. Para ello, seleccionamos **Edición, Buscar combinación** y aparecerá un cuadro de diálogo en el que debemos escribir los parámetros de la búsqueda. Aunque la sintaxis que utiliza este cuadro de diálogo es bastante sencilla, podemos utilizar el Asistente para búsquedas gramaticales, al que podemos acceder haciendo clic directamente en el botón **Asistente**. Aparecerá un nuevo cuadro de diálogo que nos preguntará **¿Sabe algún dato acerca de la palabra que está buscando?** Como sabemos que la categoría morfosintáctica, escogemos **Sé algún rasgo morfosintáctico acerca de la palabra que busco** y en las siguientes listas que nos aparecerán **Sé la categoría morfosintáctica.** y **Es un adjetivo.** Como ya hemos terminado de definir la búsqueda, elegimos por último **No, ya he terminado de definir la búsqueda que quería.**, con lo que volveremos al cuadro de diálogo anterior. Ahora basta con hacer clic en el botón **Buscar** para que aparezcan todos los resultados de nuestra búsqueda.

Como podemos observar, aparecerá en azul el número de línea de cada línea en la que se encuentre alguna de las palabras buscadas seguido de estas palabras junto con el contexto de la línea en la que aparecen. Para determinados usos, nos sería más útil disponer sencillamente de una lista con las palabras buscadas, sin necesidad de ver el contexto. Para ello, volvemos al cuadro de diálogo de búsqueda de combinaciones haciendo clic nuevamente en **Edición** y en **Buscar combinación**. Una vez en este cuadro de diálogo, activamos la casilla de verificación **Ver como lista** y hacemos clic en **Buscar**. Los resultados que aparecerán serán idénticos a los anteriores pero, en vez de ver el contexto, aparecerá sencillamente una lista en la que se muestra cada palabra junto al número de apariciones.

4. Introducción de nuevas palabras en el diccionario

Aunque el diccionario que incluye de partida PETRAtag es bastante completo, en muchos textos habrá palabras que el diccionario no incluye y, por tanto, no puedo etiquetar correctamente. Para ver una lista de todas estas palabras desconocidas, basta con hacer clic en el menú **Herramientas** y seleccionar **Estadísticas**. Aparecerá un cuadro de diálogo en el que podemos hacer clic en **Buscar palabras no etiquetadas** para que aparezca una lista con todas estas palabras. Como podemos observar, junto a cada palabra aparece un número que indica el número de veces que aparece esa palabra desconocida en el texto. Hay que tener en cuenta, que si disponemos de poco tiempo, podemos optar por introducir únicamente las palabras más frecuentes.

Para introducir una palabra, hacemos clic en el menú **Herramientas** y elegimos **Diccionarios**. Aparecerá un cuadro de diálogo con todas las palabras que incluye el diccionario. Para añadir una palabra tenemos que escribirla en el primer cuadro de texto, **Palabra:**. En el siguiente cuadro de texto, **Etiqueta:**, escribiremos su etiqueta (que podemos elegir consultando el apéndice A). Por último, en el tercer cuadro de texto, **Lema:**, si lo deseas puedes escribir el lema de la palabra. Sin embargo, también podemos optar por dejar este cuadro de texto en blanco y, de manera predeterminada, se asignará como lema de palabra el texto que hayamos escrito en el campo **Palabra:**.

5. La consola

En la parte inferior de la pantalla, debajo de la barra separadora se encuentra la consola. Aquí podemos introducir, siguiendo una sintaxis específica del programa, varios comandos que nos permiten automatizar una serie de acciones que podemos guardar en archivos de texto con el objetivo de repetir dichas acciones siempre que deseemos y modificarlas según sea necesario.

Los comandos que están disponibles actualmente son los siguientes:

CARGAR RUTAARCHIVO

Carga el archivo *RUTAARCHIVO*.

DEFINIR NOMBRELISTA CONDICIÓN

Crea una lista llamada *NOMBRELISTA* con todas las palabras del archivo cargado que cumplen la condición *CONDICIÓN*. La condición sigue la sintaxis que aparece en el cuadro **Buscar combinación**, por lo que podemos crearla utilizando el asistente que allí se incluye y pegarla en la consola.

FILTRAR NOMBRELISTA CONDICIÓN

Elimina de la lista *NOMBRELISTA* todas las palabras que no cumplen la condición *CONDICIÓN*. Esta condición se aplica al número de veces que aparece la palabra (frecuencia), por lo que suele consistir en un número máximo o mínimo.

RESTAR NOMBRELISTA1 NOMBRELISTA2

Elimina de la lista *NOMBRELISTA1* todas las palabras que aparecen en la lista *NOMBRELISTA2*.

MOSTRAR NOMBRELISTA

Muestra en la pantalla el contenido de la lista *NOMBRELISTA*.

Como ejemplo, la siguiente secuencia de comandos muestra todas las palabras que aparecen más de diez veces en el archivo *Texto1* y no aparecen en el archivo *Texto2*.

```
CARGAR C:\Textos\Texto1.txt
DEFINIR LISTA1 0<E1()>
CARGAR C:\Textos\Texto2.txt
DEFINIR LISTA2 0<E1()>
FILTRARFRECUENCIA LISTA1 >10
RESTAR LISTA1 LISTA2
MOSTRAR LISTA1
```

6. Contacto

PETRAtag se encuentra en las primeras fases de desarrollo, por lo que agradeceríamos cualquier duda o sugerencia que nos haga llegar. También agradeceríamos que nos

informáseis de cualquier error que haya aparecido. Para todo ello, nuestra dirección de contacto es: tag@equus-trad.com

7. Bibliografía

M. Civit (2003) Criterios de etiquetación y desambiguación morfosintáctica de corpus en español Sociedad Española para el Procesamiento del Lenguaje Natural, Colección Monografías, número 3 ISBN: 84-600-9944-X

Apéndice A: Guía de etiquetado

1. Adjetivos

Adjetivos			
Pos.	Atributo	Valor	Código
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
3	Grado	Apreciativo	A
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Caso	-	0
7	Función	Participio	P

2. Adverbios

Adverbios			
Pos.	Atributo	Valor	Código
1	Categoría	Adverbio	R
2	Tipo	General	G
3	-	-	0
4	-	-	0
5	-	-	0

3. Artículos

Artículos			
Pos.	Atributo	Valor	Código
1	Categoría	Artículo	T
2	Tipo	Definido	D
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
5	Caso	-	0

4. Determinantes

Determinantes			
Pos.	Atributo	Valor	Código
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Caso	-	0
7	Poseedor	1ª persona-sg	1
		2ª persona-sg	2
		3ª persona	0
		1ª persona-pl	4
		2ª persona-pl	5

5. Nombres

Nombres			
Pos.	Atributo	Valor	Código
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	-	0
6	Género semántico	-	0
7	Grado	Apreciativo	A

6. Verbos

Verbos			
Pos.	Atributo	Valor	Código
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Condicional	C
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P

		Imperfecto	I
		Futuro	F
		Pasado	S
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

7. Pronombres

Pronombres			
Pos.	Atributo	Valor	Código
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	1ª persona-sg	1
		2ª persona-sg	2
		3ª persona	0
		1ª persona-pl	4
		2ª persona-pl	5
8	Politeness	Polite	P

8. Conjunciones

Conjunciones			
Pos.	Atributo	Valor	Código
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S
3	-	-	0
4	-	-	0

9. Numerales

Numerales			
-----------	--	--	--

Pos.	Atributo	Valor	Código
1	Categoría	Numeral	M
2	Tipo	Cardinal	C
		Ordinal	O
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
5	Caso	-	0
6	Función	Pronominal	P
		Determinante	D
		Adjetivo	A

10. Interjecciones

Interjecciones			
Pos.	Atributo	Valor	Código
1	Categoría	Interjección	I

11. Abreviaturas

Abreviaturas			
Pos.	Atributo	Valor	Código
1	Categoría	Abreviatura	Y

12. Preposiciones

Preposiciones			
Pos.	Atributo	Valor	Código
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
3	Género	Masculino	M
4	Número	Singular	S

13. Signos de puntuación

Signos de puntuación			
Pos.	Atributo	Valor	Código
1	Categoría	Puntuación	F