UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Escuela de Ingeniería Informática

# AMPLIACIÓN DE LA PARAMETRIZACIÓN MORFOLÓGICA DE TEXTOS

**Autor:** Jorge Bueno Godoy

**Tutor:** D. Francisco Javier Carreras Riudavets

**Cotutor:** D. Zenón Hernández Figueroa

**Titulación:** Ingeniería Informática

**Fecha:** 25/06/2014

Proyecto fin de carrera de la Facultad de Informática de la Universidad de Las Palmas de Gran Canaria presentado por el alumno:

**Jorge Bueno Godoy**

**Título del proyecto:** Ampliación de la parametrización morfológica de textos

**Tutor:** D. Francisco Javier Carreras Riudavets

**Cotutor:** D. Zenón Hernández Figueroa

A mi madre, hermana y Jessica.

# Agradecimientos

En primer lugar quiero dar mi agradecimiento a las dos personas que me han acompañado a lo largo de este proyecto. A mi madre, por darme la oportunidad que no pudo tener y poner a mi alcance todos los medios necesario para convertirme en la persona que soy hoy. A Jessica, por ser fuente de cariño y comprensión, especialmente durante la realización de este proyecto.

Del mismo modo quiero agradecer a Paco, mi tutor, su gran trabajo e implicación, por atender mis dudas y concederme más tiempo del estrictamente necesario para resolverlas.

Para finalizar quiero dar las gracias a mi hermana, familia, amigos, colegas, profesores y a todos aquellos que me han apoyado tanto a mí como a este proyecto.

# ÍNDICE

# 1 Introducción

## *1.1 Parametrización morfológica de textos*

Al igual que el resto de especies de nuestro planeta, el ser humano es capaz de comunicarse entre sí, sin embargo, a diferencia ellos nuestra capacidad ha ido mucho más allá del simple hecho de informar. Casi al mismo nivel que el concepto que transmiten en sí mismo, la forma en la que las palabras llegan al receptor es capaz de facilitar su comprensión e incluso la aceptación de las mismas. Cuando hablamos del lenguaje escrito dicha afirmación cobra una importancia especial.

A diferencia de lo que ocurre en la comunicación oral, cuando pretendemos hacernos entender mediante la escritura, estamos exigiendo a nuestros lectores que empleen tiempo extra y les privamos de la información procedente del lenguaje corporal. Por su lado, la falta de entonación puede llevarnos a confundir totalmente el objeto del mensaje y, en consonancia con lo anterior, a resultar una completa pérdida de tiempo si no somos capaces de entenderlo, o peor aún, si nos hace creer que lo hemos entendido cuando no ha sido así.

Independientemente del género, todo escrito debe cumplir con ciertos requisitos para facilitar su comprensión: coherencia, cohesión, comprensión, intencionalidad, morfología o la sintaxis son algunos ejemplos. La mayoría de estas características implica la previa lectura del texto para poder determinar si finalmente se trata o no de un texto de calidad, ya que la mayor parte de estos requisitos son factores no cuantitativos.

A lo largo de la historia han aparecido personas con un don para la capacidad de transmitir mediante la escritura, desde los clásicos Miguel de Cervantes, Lope de Vega o Francisco de Quevedo hasta otros más actuales como García Lorca, Miguel Delibes o Camilo José Cela pasando por una considerable cantidad de ilustres de la palabra. Sería imposible copiar su arte, pero no lo es comparar la estructura y morfología de los textos de las nuevas generaciones con los suyos para, al menos, hacernos una idea de hasta qué punto se parecen sus formas (aunque no sus contenidos).

Como si de la otra cara de la moneda se tratara, el avance de la tecnología nos ha proporcionado instrumentos de comunicación cada vez más rápidos y manejables, así como la capacidad de escribir casi cualquier cosa en cualquier momento y que sea accesible para cualquier persona; hecha la ley, hecha la trampa. La aparición de los SMS, que requerían más tiempo para ser escritos y su coste inicial fomentó la aparición de pseudolenguajes orientados a ahorrar pulsaciones y espacio pero a dificultar su comprensión, entre otras razones debido a su falta de elementos de puntuación, tildes, el abuso de abreviaciones y la carencia de estructuración en general. Cual virus se extendió hasta los rincones de la red de redes y a día de hoy, aun cuando son en su mayor parte innecesarios debido a los nuevos teclados en

dispositivos portátiles y al abaratamiento y las distintas opciones de comunicación, siguen haciendo gala de presencia.

Por todo ello se ha ampliado el software capaz de generar estadísticas completas de un texto basándose únicamente en sus aspectos morfológicos (sustantivos, verbos, adjetivos, artículos, determinantes, pronombres, etc.), de forma que se puedan contrastar los textos de aquellas personas que lo deseen con obras literarias de personalidades de renombre dentro del género, proporcionando así una comparación mediante la denominada *Parametrización Morfológica* del texto. Así mismo se ha añadido información sobre el vocabulario que contiene con elementos como si la palabra está contenida en la RAE o si está en desuso.

## 1.2 Estado actual del tema

A día de hoy sigue sin existir una alternativa a la herramienta Paramtext, la cual sigue acumulando herramientas que la hacen cada vez más funcional. Así mismo, el terreno ganado por redes sociales y publicaciones online frente a sus homónimos fuera del entorno virtual potencia aún más si cabe la posibilidades de esta aplicación. Su integración en ellas puede ser la clave para un éxito a gran nivel.

## 1.3 Estructuración de la memoria

La presente memoria documenta las distintas etapas que han sido necesarias para desarrollar este proyecto. En los primeros capítulos se abordan los objetivos, la metodología empleada y la planificación temporal. Los siguientes capítulos han sido divididos en las distintas fases del proyecto según la ingeniería del software.

Junto a esta memoria se proporciona un DVD que contiene lo siguiente:

- **Carpeta memorias:** que incluye el presente documento y su antecesor, el creado por Juan Carlos Santana Herrera en el proyecto inicial.
- **Carpeta Paramtext 2.0:** contiene el código del programa y los archivos necesarios para probar la aplicación.
- **Carpeta PFC:** donde se encuentran los documentos PFC entregados en administración.

# 2 Objetivos

En primer lugar y de forma directa, la aplicación final debe generar una comparación mediante la *Parametrización Morfológica* de un texto en base a otro de calidad reconocida, con lo que se pretende que el usuario pueda saber si el texto que ha escrito o pretende leer tiene calidad morfológica comparativamente con aquel otro que haya utilizado como referente. De igual modo de incluir información adicional acerca del vocabulario utilizado que pueda ser relevante a ahora de determinar si un texto es de fácil comprensión.

En segundo lugar y de forma potencial, esta herramienta podría servir como base para futuras ampliaciones y aplicaciones en la red, mediante las cuales se pudiera saber de antemano, en base a unos patrones acordados, si la calidad de la información a la que vamos a acceder es, cuanto menos, aceptable. Esto nos permitiría el uso de filtros a la hora de acceder a los contenidos que nos interesan, desde foros hasta buscadores (como Google) pasando por redes sociales (como Facebook) y distribuidoras de información (como periódicos o diarios online) podrían hacer uso de ella para asignar o comprobar los niveles de calidad de sus publicaciones. De manera similar a los sellos de certificación de calidad que se usan en alimentos u otros productos, dicha herramienta podría llegar a convertirse en un signo de calidad lingüística.

Para ello han cumplido los siguientes objetivos:

- Realizar un estudio del programa original para comprender su metodología y estructura, de tal forma que las modificaciones que se han realizado mantengan la misma línea de funcionamiento y permita añadir aún más funciones en el futuro sin tener que hacer una ingeniería inversa del original y otra completamente distinta de su primera ampliación.
- Diseñar y desarrollar la ampliación de la interfaz para mostrar la nueva información y la comparación de la misma de forma clara, manteniendo la compatibilidad y el estilo con la creada previamente.
- Definir un modelo de software aplicando los criterios aprendidos durante la carrera. Haciendo especial hincapié en la programación modular y los comentarios para facilitar la reutilización del código y sus futuras ampliaciones.

# 3 Metodología

La metodología para el desarrollo de software es un marco de trabajo usado para estructurar, planificar y controlar el proceso de desarrollo en sistemas de información. No estipular la metodología adecuada puede generar problemas a lo largo de la vida del software, sobre todo cuanto más largos y complejos sean.

## *3.1 Ciclo de vida en espiral*

La metodología elegida para la realización de este proyecto ha sido la misma que para su antecesor, el ciclo de vida en espiral. Su facilidad para incluir nuevas funcionalidades a lo largo del tiempo lo hace un candidato excelente para proyectos tan grandes y ambiciosos como lo el Paramtext. De esta forma es fácil llevar un control por etapas sobre el proyecto, lo que evita que errores iniciales puedan ser arrastrados hasta las etapas finales.

En cada iteración del ciclo se dan las siguientes fases del software:

1 **Determinar los objetivos:** Al hacerse en cada iteración permite definir los objetivos en función del software que ya se ha generado, con lo que no es necesario abordar aquellos que son independientes o que pudieran ser más fácilmente establecidos con posterioridad en base al estado del programa.
2 **Análisis del riesgo:** Se lleva a cabo un análisis para cada uno de los riesgos identificados del proyecto y se definen los pasos a seguir para reducirlos. Su mayor peligro es no ver más allá de la iteración actual cuando esta pueda estar relacionada con otras anteriores o posteriores.
3 **Desarrollar y probar:** Se desarrolla el software y se valida.
4 **Planificación:** Se revisa el estado actual del proyecto, permitiendo avanzar si se determina que funciona como debería o volver a una iteración anterior. Hacer esto en cada iteración es una gran ventaja que nos permite no continuar trabajando sobre algo que no es válido, y que de no haberse estudiado en este momento podría invalidar mucho trabajo posterior que se base en él.

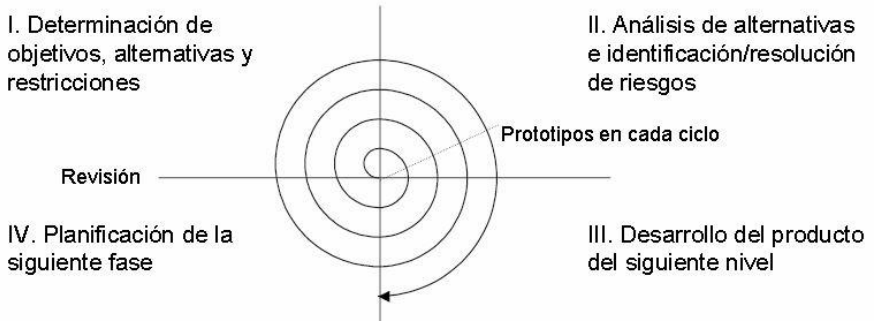



*Figura 3.1*

# 4 Recursos necesarios

En este capítulo se detallan los recursos hardware y software que han sido utilizados para llevar a cabo la realización del proyecto.

## *4.1 Hardware*

Puesto que para la realización de este tipo de proyecto no es necesario utilizar un hardware específico, el alumno ha hecho uso únicamente de los recursos propios con los que contaba. Cabe destacar la necesidad de conexión a internet para hacer uso de los servicios en línea necesarios para el correcto funcionamiento de Paramtext.


**Ordenador personal**

- Intel i7 2600
- 8 GB RAM
- 1 TB de disco duro
- Conexión a internet 30 Mb


**Memoria USB**

- 32 GB

## 4.2 Hardware Virtual

Para compartir el proyecto se optó por usar un servicio de almacenamiento virtual, con lo que no era necesario el envío directo de información mediante correo electrónico ni Pendrivers.


**Dropbox**

- +800 MB para las distintas versiones de Paramtext y aplicaciones auxiliares.
- ~300 MB para los archivos de la RAE

## 4.3 Software

Dado que el proyecto de ampliación pretendía respetar el diseño y la tecnología usada en el proyecto original no se han utilizado herramientas ni tecnologías adicionales.

| RECURSO | PRECIO |
|---|---|
| XHTML | Gratuito |
| XML | Gratuito |
| CSS | Gratuito |
| Java Script | Gratuito |
| AJAX | Gratuito |
| UML | Gratuito |
| Microsoft.NET | Gratuito |
| C# | Gratuito |
| Microsoft Visual Studio 2010 | Gratuito – Licencia MSDN AA |
| ASP.NET AJAX Control Toolkit | Gratuito |
| Microsoft Chart Control | Gratuito |
| Bytescout Document SDK | Gratuito |
| PDFBox | Gratuito |
| Star UML | Gratuito |

*Tabla 4.1*

*Puede encontrarse una descripción detallada de ellas en la memoria original de Paramtext (Pág. 8 de la memoria/Pág. 20 del documento PDF; Apartados 4 y 5)*

# 5 Plan de trabajo y presupuesto

En este capítulo se detallan las distintas etapas en las que se dividió el proyecto y una estimación del tiempo necesario para llevarlo a cabo.

## *5.1 Fases del proyecto*

Se decidió organizar el proyecto en 5 fases. Hay que recordar que el ciclo de vida en espiral genera múltiples iteraciones, por lo que dichas fases se dieron lugar tantas veces como iteraciones fueron necesarias en la ejecución del proyecto:

1. **Documentación:**
   Estudio de la estructura y funcionamiento de Paramtext para poder realizar la ampliación respetando el estilo y código original en la medida de lo posible. Además se incluye el tiempo requerido para la elaboración del actual documento.

2. **Análisis:**
   Se estudiaron los requerimientos funcionales y no funcionales del sistema para entender con exactitud la naturaleza de los programas necesarios para realizar la ampliación mientras se tiene en cuenta la información obtenida en la fase de Documentación

3. **Diseño:**
   En esta fase se determinó la estructura que debían seguir las modificaciones para evitar entrar en conflicto con la versión original, respetando su código e integrándose de forma amigable.

4. **Desarrollo:**
   Implementación de las ampliaciones dentro del entorno Paramtext mediante la utilización de las mismas herramientas originales.

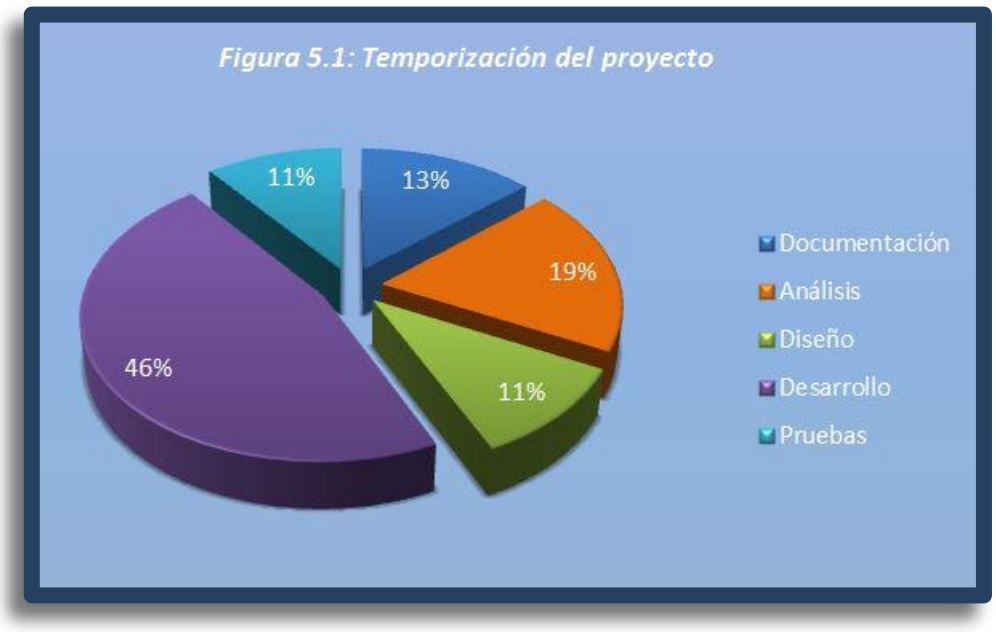
5. **Pruebas:**
   Selección y realización de pruebas para garantizar que la integración de las modificaciones y ampliaciones no dio lugar a conflictos.

## 5.2 Temporización del proyecto

A continuación se muestra la estimación del tiempo necesario para realizar cada una de las fases.

| FASE | HORAS |
|---|---|
| **Documentación** | 160 |
| **Análisis** | 225 |
| **Diseño** | 125 |
| **Desarrollo** | 550 |
| **Pruebas** | 125 |
| **Total** | **1185** |

*Tabla 5.1: Temporización del proyecto*



## *5.3 Presupuesto del proyecto*

En este apartado se elabora el presupuesto para la realización del proyecto:

**Costes laborales:**

Suponiendo que se trabaja para una empresa y que el alumno cobra 9€ la hora y el tutor (personal con más experiencia) 18€ la hora, los costes laborales son:

- Alumno: 9€/hora x 550 horas (implementación) = 4950€
- Tutor: 18€hora x 50 horas = 900 horas

**Costes materiales:**

En cuanto a los costes materiales hay que considerar su vida útil y que no sólo se utilizan para realizar el proyecto. Considerando que el material será utilizado un 50% del tiempo en cuanto al hardware, un 10% de la capacidad de internet para el proyecto y el resto del tiempo y capacidad para otras actividades, y que la duración del proyecto se establece en 8 meses. El coste de los materiales necesario será:

- Coste del ordenador: (800€ x 0,5 uso x 8 meses) / 48 meses = 62,5€
- Cable 30 MB: 45€ x 0,1 uso x 8 meses = 36€

**Costes de material fungible:**

Son los gastos generados por el uso diario del material (papel, tinta, etc.): 60€

**Coste de documentación:**

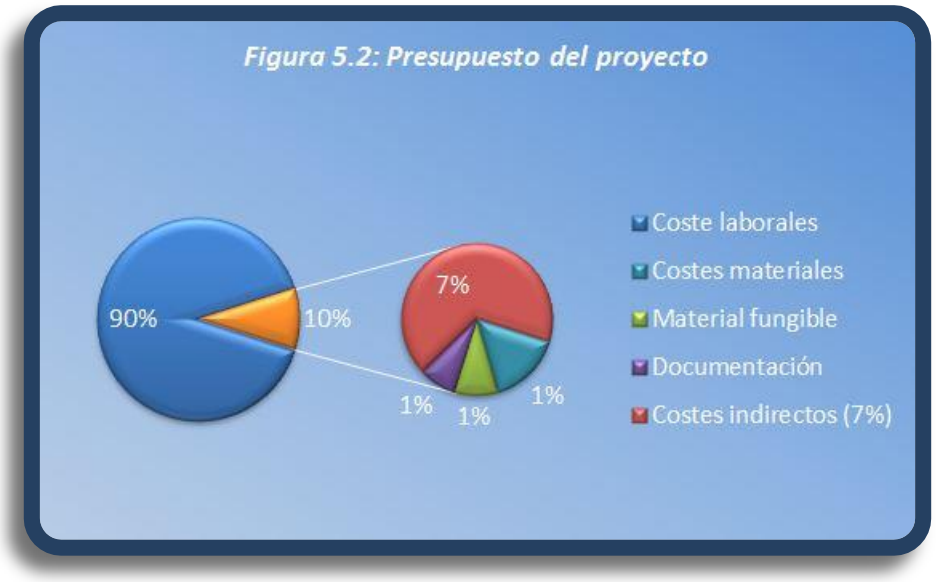
Son los gastos necesarios para la documentación: 50€

**Costes indirectos:**

Son los costes de las cosas que no están vinculadas directamente al PFC, pero que son necesarias para que el PFC se desarrolle. Se considera que el 8% del resto de gastos corresponde a gastos indirectos.

| CONCEPTO | COSTE |
|---|---|
| **Coste laborales** | 5850 |
| **Costes materiales** | 98,5 |
| **Material fungible** | 60 |
| **Documentación** | 50 |
| **Subtotal** | **6058** |
| **Costes indirectos (7%)** | 424,1 |
| **Total** | **6482,1** |

*Tabla 5.2: Presupuesto del proyecto.*



# 6 Estudio del dominio del problema

Este proyecto nace de la necesidad de mejorar la funcionalidad de Paramtext y de dar sentido a unos valores numéricos asociados a las obras de los usuarios que inicialmente pueden ser difíciles de interpretar. Para tal empresa se ha optado, por un lado, en ofrecer la capacidad de comparar los resultados obtenidos con aquellos que provienen de obras de renombre, de esta forma el usuario podrá saber en qué medida se acercan o alejan sus valores de aquellos que se obtienen en obras de calidad reconocida. Por otro lado se ha añadido información adicional acerca del vocabulario, lo que facilita el valorar la "calidad" de las palabras usadas en el texto y no sólo su estructura.

## *6.1 La estructura de un texto mediante valores*

¿Qué representa un treinta en el número de párrafos en un texto de mil palabras más allá de que hay mil palabras distribuidas a lo largo de treinta párrafos? ¿Es mucho o poco? ¿Significa lo mismo en un texto de poesía que en otro escrito en prosa? ¿En una novela de acción que en una de romance?

La interpretación de valores numéricos cuando se habla de textos no es sencilla y una buena forma de entenderlos es viendo cuales son los que obtiene otra obra en comparación. Por ejemplo, las novelas de acción suelen usar una gran cantidad de frases cortas y sencillas para transmitir al lector la sensación de dinamismo y estrés a la que están sometidos los personajes en las escenas críticas, pero el uso excesivo de dicha técnica es contraproducente pues las frases cortas y simples le quitan riqueza al lenguaje y a la estructura del texto, hasta hacerlo parecer el texto de un niño que no sabe elaborar frases más complejas. Si permitimos al usuario comparar su texto de acción con el de otro que haya sido un referente en el género estaremos ayudándole a decidir hasta qué punto quiere acercarse o no a ese estilo de escritura ¿Quiere estar dentro de la norma y usar un número similar de frases cortas y de longitud parecida o por el contrario prefiere innovar en el género y hacer todo lo contrario? No será Paramtext quien le diga lo que tiene que hacer sino quien le informe de lo que está haciendo, por si eso es lo que pretendía o no.

Dicha comparación es en sí misma complicada, debido al gran número de variables que entran en juego, desde el tamaño de los párrafos y las frases hasta el número de las segundas dentro de los primeros y la cantidad de los primeros en el total de la obra, etcétera, o mejor dicho, un larguísimo etc. Para facilitarlo, Paramtext se reorganiza y muestra los valores de ambas obras en las mismas tablas y gráficas delimitándolas con claridad, de forma que sea mucho más sencillo entender los datos que nos ofrece.

## 6.2 La importancia de usar un vocabulario adecuado

Ya hemos hablado de la importancia de la estructura dentro de un texto, pero hay otros factores que pueden acabar con una buena obra antes de poder sumergirse en ella, el vocabulario.

No hay que ser un genio para saber que un libro infantil debería usar palabras sencillas mientras que otro de intrigas políticas en la edad media podría hacer uso de palabras en desuso actualmente y una sintaxis más complicada. Paramtext ha incluido una gran cantidad de información adicional en lo que se refiere al vocabulario, entre las que se encuentran si una palabra aparece en la RAE, está en desuso actualmente o el número de significados distintos que tiene. De esta manera el escritor puede hacerse una idea del tipo de vocabulario que está usando y determinar si ese es el camino que quiere seguir. Para llevar a cabo esta tarea fue necesario extraer una gran cantidad de información del diccionario de la RAE, que contiene alrededor de cien mil palabras.

No menos importante es tener conocimiento de la forma en la que construimos las frases y no sólo de las palabras individuales que usamos en ellas. Repetir constantemente cosas como “Por lo tanto” o el uso de “coletillas” es un problema común y para ayudar a solucionarlo se ha incluido una nueva tabla de n-gramas, en donde se pueden consultar todos los existentes con tamaños comprendidos entre dos y cinco, indica cuantas veces se ha repetido, donde pareció por primera vez y el centro de gravedad, que nos ayuda a determinar en qué zona ha sido más utilizada. Además de localizar aquellas “coletillas” propias, usado en el modo comparación podemos determinar en cuanto se parecen nuestras formas a las de otro autor.

## 6.3 Información incluida en el vocabulario

Los datos seleccionados para ser añadidos fueron los siguientes:

- **Categoría gramatical:**
  Clasificación de las palabras según su naturaleza morfológica, por ejemplo, sustantivo, pronombre, verbo, etc.

- **Etimologías:**
  Determina el número de fuentes de las que proviene la palabra.

- **Acepciones:**
  Indica el número de significados distintos que tiene una misma palabra

- **Posición de la acepción:**
  Muestra la posición de la categoría gramatical de la palabra dentro del conjunto de significados distintos que se muestran en la RAE. Si una palabra tiene 5 significados distintos donde funciona como pronombre en los tres primeros (1, 2 y 3) y determinante en los dos últimos (4 y 5) se mostrará un uno si la palabra actúa como pronombre y un 4 si lo hace como determinante.

- **Entradas de la acepción**:
  Contabiliza el número de entradas distintas que tiene la categoría gramatical que ha sido asignada a esa palabra, de forma que en el ejemplo anterior, indicaría un tres si fuera un pronombre porque hay tres significados distintos en los que la palabra funciona como pronombre y un dos si fuera un determinante (porque nos encontramos con dos significados distintos cuando hace de determinante).

- **Antigua:**
  Muestra si la palabra está en desuso según la decimotercera edición de la RAE

- **Aparece en la RAE:**
  Como su propio nombre indica nos informa sobre si es una palabra reconocida por la RAE o si por el contrario no pertenece realmente al castellano.

Dado que estas herramientas muestran factores tales como si se usan demasiadas palabras no reconocidas por la RAE, si están en desuso o si se usan los significados de las palabras menos utilizados en lugar de los que parecen en los primeros lugares del diccionario; en su conjunto pueden ayudar a determinar si el vocabulario utilizado era el que se pretendía o si está realmente orientado al público al que va dirigido el texto en cuestión.

## *6.4 Adquirir la información*

Para poder añadir la nueva información era necesario extraerla del lugar más fiable, por tanto se usó la página web oficial de la RAE y para hacerlo fueron necesarias tres herramientas independientes. La primera de ellas realizaba peticiones automatizadas de información y era externa a este proyecto mientras que las otras dos son explicadas en detalle en el siguiente capítulo de esta memoria.

## *6.5 La función de los n-gramas*

Los n-gramas son todos aquellos conjuntos de palabras que se pueden encontrar en un texto agrupados según su tamaño. Localizándolos y agrupándolos es posible saber si existen alguna composición de palabras que se repita demasiado a lo largo de un texto, o si dos textos distintos se parecen demasiado en sus formas independientemente de que se hayan cambiado de orden las frases, los párrafos o los capítulos.

Los n-gramas de tamaño tres (por ejemplo) serán todos los grupos de tres palabras que se encuentren en el texto, y si los buscamos en la siguiente frase el resultado sería:

Frase: “Veamos cuantos n-gramas de tamaño tres hay en esta frase”

1 Veamos cuantos n-gramas
2 cuantos n-gramas de
3 n-gramas de tamaño
4 de tamaño tres
5 tamaño tres hay
6 tres hay en
7 hay en esta
8 en esta frase

Desde los más simples de tamaño dos como podría ser “por consiguiente” a otros más largos como “dicho sea de paso”, el estudio de los n-gramas nos puede ayudar, entre otras cosas, a no pecar de repetitivos en cuanto al uso de ese tipo de expresiones.

# 7 Herramientas auxiliares

Toda la información añadida a la sección de vocabulario fue extraída de la página web de la RAE. Dado que existen alrededor de cien mil palabras (100.000) la extracción y clasificación de la información no fue un trabajo trivial, sobre todo teniendo su heterogeneidad.

La primera herramienta necesaria tenía como función realizar peticiones a la web y almacenar los resultados "brutos", sin tener en cuenta si lo devuelto era un error de conexión o la información de la palabra. Dicha herramienta era externa al proyecto por lo que no será explicada en mayor profundidad, las otras dos se detallan a continuación.

**Nota importante:** El uso de antivirus mientras se ejecutan las herramientas puede aumentar el tiempo requerido entre treinta y ochenta veces (según el antivirus) puesto que el programa debe abrir, cerrar y modificar los archivos de texto constantemente y el antivirus se dedica a inspeccionar cada una de las líneas que se modifican o añaden. Por tanto se recomienda encarecidamente desactivarlos mientras se ejecutan.

## *7.1 Limpiador WebRae*

La segunda herramienta tenía dos funciones principales, por un lado la de eliminar todo el código HTML innecesario, por lo que recorría el archivo asegurándose de eliminar imágenes, iconos y toda la información que no era relevante. La otra función consistía en encontrar aquellas palabras que no habían sido correctamente extraídas, teniendo en cuenta la razón del problema y generando archivos de información al respecto.

Los distintos problemas detectados por Limpiador WebRae son:

1. **Aviso:** No se ha podido procesar su información. Puede intentarlo más tarde
2. **Error:** Detectar un Elema duplicado con distinto índice, lo que significaba que se había extraído dos veces la misma palabra pero se había contabilizado como si fuera la siguiente en la lista.
3. **Error:** Número repetido y vacío, indicaba que se había intentado extraer la misma entrada más de una vez y el resultado había sido nulo.
4. **Error:** Palabra con información incompleta; se daba cuando la información extraída de la RAE no disponía de toda la información que supuestamente debía tener la palabra, debido a la heterogeneidad del código usado en la RAE la detección de estos errores fue especialmente difícil de configurar, pues no siempre faltaba la misma información.
5. **Error:** Error de conexión; simplemente se había producido un error en la conexión

Además de la detección de avisos y errores la herramienta presentaba las siguientes propiedades:

- Generaba un único archivo de salida a partir de uno o varios archivos de entrada, pudiendo barrer un directorio completo en busca de archivos con el formato "RAE*.html".
- Podía añadir al mismo archivo de salida nueva información en sucesivas ejecuciones del programa, de tal manera que el archivo de salida se podía ir actualizando a medida que se extraía nueva información de la página de la RAE.
- Creaba archivos "log" donde se enumeraban los distintos avisos y errores, así como el número de entrada que lo había generado

Debido al gran tamaño de los archivos con los que se iba a trabajar se optó por informar al usuario de la cantidad de archivos que han sido limpiados, los que quedan por limpiar y el tamaño de dichos archivos; además de mostrar una barra de progreso real que muestra la cantidad de trabajo restante en base al volumen de información que queda por tratar y no en base al número de archivos. Por ejemplo, si se fueran a limpiar tres archivos, uno de dos (2) megas, otro de cinco (5) megas y un tercero de tres (3) megas (10 megas en total) la barra se situaría en el 20% tras acabar el primer fichero (2 de 10 megas) luego al 70% (7 de 10 megas) y por último se completaría al finalizar el tercer fichero.

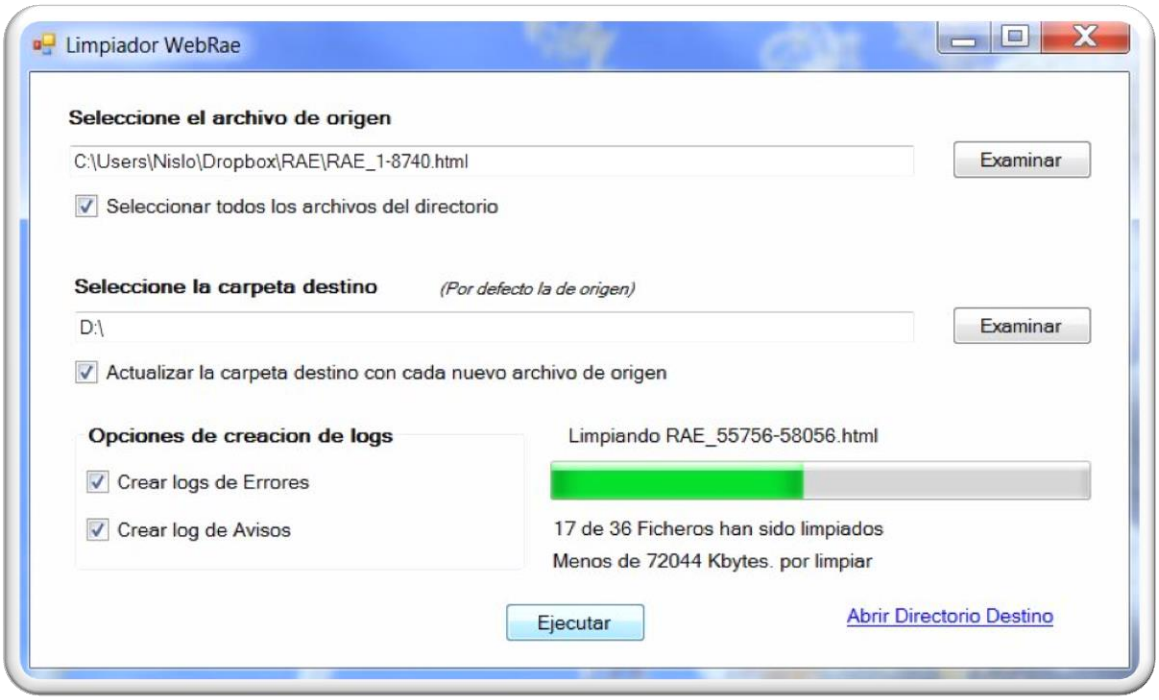



*Figura 7.1*

## 7.2 Extractor ArchivoLimpioRae

La tercera herramienta era la encargada de buscar entre todo el código restante de las páginas web de la rae la información relevante para cada palabra y guardarla en dos archivos. El primero contenía la relación entre la "key" (número de entrada en la RAE para esa palabra) la propia palabra y su etimología (de donde proviene) con el siguiente formato (la información separada por arrobas):


*Figura 7.2*

Como se puede observar, el código original estaba repleto de etiquetas innecesarias o duplicadas y ni siquiera mantenía el mismo formato a lo largo de las palabras con lo que el proceso de producción de esta herramienta fue mucho más largo y complicado de lo deseado.

El segundo archivo relacionaba la "key" de una palabra con el número de etimologías, las veces que aparecía cada una y en qué lugar estaba situada de entre todas ellas (considerando que las que estuvieran antes en el diccionario fueran las más relevantes o utilizadas). El formato utilizado fue el siguiente (información separada por arrobas)

**53**@1@1@1@Verbo transitivo@

**53**@0@1@1@También como verbo pronominal@

**54**@1@1@1@Nombre masculino@

**55**@10@1@6@Verbo transitivo@

**55**@10@7@4@Verbo pronominal@

**55**@0@5@2@Más como verbo pronominal@

**56**@1@1@1@Nombre masculino@

57@2@1@2@Adjetivo@

57@0@2@1@También como sustantivo@

58@3@1@3@Nombre masculino@

*Figura 7.3*

El problema creció al encontrarnos con acepciones que no estaban situadas al inicio de la palabra sino al final de ella con los formatos “usada también como”, “usada menos como” y “usada más como”. Esta última manera de llamarla nos sigue resultado incomprensible a día de hoy, pues si una palabra (como el caso 55 del ejemplo de la figura 7.3) “se usa más como verbo pronominal” ¿Por qué no clasificarla como verbo pronominal y poner “usada menos como verbo transitivo” para darle más importancia a aquella acepción que más se usa?

Adicionalmente existe una opción en el programa para añadir el resto de la información de cada palabra en este último fichero de texto por si fuera relevante en el futuro. Dicha información no está tratada y se corresponde en su mayoría con las distintas locuciones que pueda tener cada palabra, a modo de curiosidad, la palabra “agua” tiene más de cien locuciones distintas.

Por último, y debido a que este proceso es aún más largo que el primero de limpieza, el programa también presenta una barra de progreso basado en el número de palabras que quedan por extraer, lo que facilita comprobar su estado y el tiempo estimado.

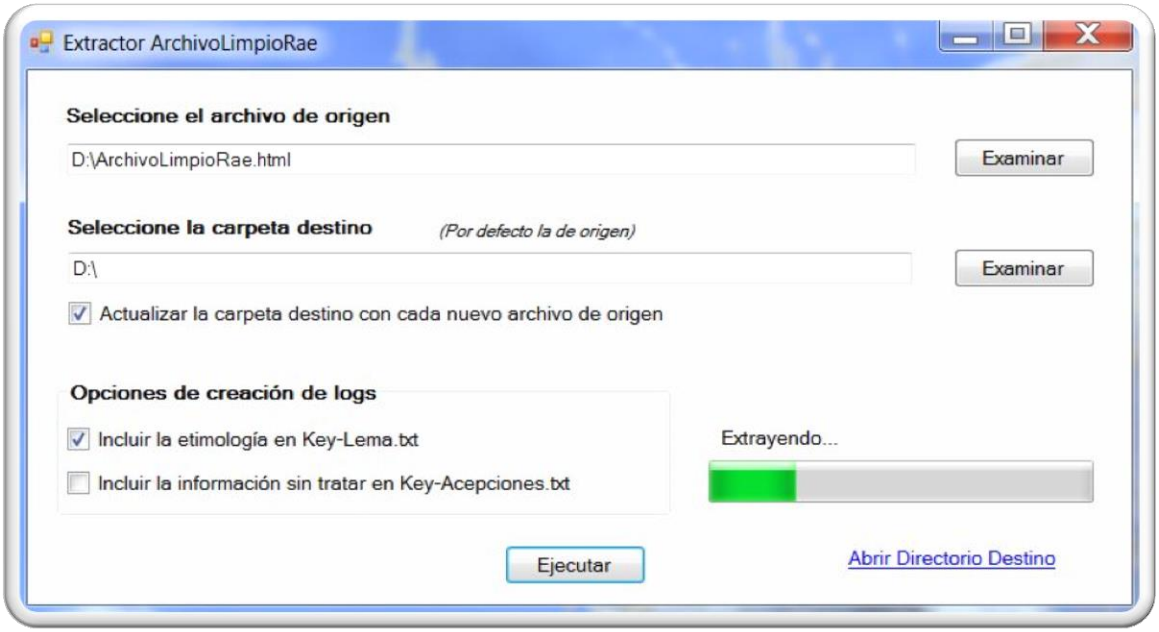

Extractor ArchivoLimpioRae
Seleccione el archivo de origen
D:\ArchivoLimpioRae.html
Examinar
Seleccione la carpeta destino
(Por defecto la de origen)
D:\
Examinar
Actualizar la carpeta destino con cada nuevo archivo de origen
Opciones de creación de logs
Incluir la etimología en Key-Lema.txt
Incluir la información sin tratar en Key-Acepciones.txt
Extrayendo...
Ejecutar
Abrir Directorio Destino

*Figura 7.4*

# 8 Análisis

La primera fase de todo proyecto es la de su análisis, siendo la base sobre la que se sustentará el resto del mismo. De nada vale un perfecto diseño y una programación impecable si no se satisfacen las necesidades de los usuarios y es por eso que existen los requisitos del software, que son todos aquellos requerimientos que el usuario impone a un software para que se justifique su uso. Por supuesto no sólo habrá que tener en cuenta los requisitos funcionales (que el programa haga lo que se pide) sino también aquellos no funcionales (que el programa lo haga de la forma adecuada, en forma y seguridad por ejemplo).

Ha sido parte importante a lo largo de todo el proyecto la idea de mantener un código homogéneo dentro de lo posible, evitando que posibles modificaciones posteriores necesitaran realizar ingeniería inversa sobre Paramtext y su ampliación por separado. Cabe destacar que el código original estaba muy bien comentado y su gran modularidad ha sido de ayuda a la hora de incluir nuevas funcionalidades. Aun y así es imposible adaptar la totalidad del proyecto a los patrones seguidos en su concepción inicial y algunas estructuras y módulos han necesitado ser ampliados, modificados o incluso eliminados y vueltos a hacer.

## *8.1 Requisitos del software*

A groso modo lo que se pretende ofrecer al cliente es la posibilidad de comparar su texto con otro que tenga una calidad reconocida, de forma que los valores que antes mostraba Paramtext cobren sentido más allá de un simple número. Dicha comparación debe ser fácil y rápida de realizar, o sería igual de productivo abrir el programa en dos navegadores distintos y poner uno al lado de otro, por tanto, este proyecto no consistía en hacer “copiar y pegar” al lado de la tabla o gráfica existente sino que proporciona una combinación y reestructuración de datos tanto a nivel externo como interno.

Adicionalmente el cliente quiere incluir información adicional al vocabulario, una parte de ella proveniente de la RAE y otra, los n-gramas, extraídos del propio texto mediante un nuevo algoritmo. Además se pretende utilizar el servicio silabeador del grupo TIP[1] en lugar del código incrustado del que hacía uso, de este modo cualquier mejora en el servicio se reflejaría en el Paramtext sin necesidad de tener que modificar el programa.

[1] Text & Information Processing (TIP), grupo formado por profesores del Departamento de Informática y Sistemas de la Universidad de Las Palmas de Gran Canaria dedicados a la investigación y desarrollo de herramientas y aplicaciones en áreas de la lingüística computacional, lexicografía, procesamiento del lenguaje natural, indización, almacenamiento de información y conocimiento y educación tecnológica. (http://tip.dis.ulpgc.es/)

Con vistas a posibles ampliaciones futuras, la información adicional proveniente de la RAE se estructurará de forma que pueda usarse para presentar información de carácter subjetivo basada en ella. Para realizar la tarea de extracción de información fue necesaria la utilización de tres herramientas adicionales, dos de ellas explicadas en el capítulo anterior y que fueron creadas también como parte del proyecto.

Entrando en las funciones de comparación, se prestará especial interés en permitir que los administradores del programa sean capaces de ampliar el abanico de texto ofrecidos para comparar sin necesidad de tocar ni una sola línea de código. Mediante el uso de archivos XML y TXT externos al programa es posible añadir, modificar, eliminar obras, colecciones y la información que se muestra de las mismas. Esto permitirá además, que dichas modificaciones se den sin necesidad de detener el programa en ningún momento y se producirán en tiempo real, de esta forma no será necesario dejar de ofrecer el servicio ni tener que reiniciarlo para que los usuarios puedan ver los cambios.

Así mismo se requiere que el programa permita hacer la comparación desde su propio apartado, pero también que exista la opción de deshacer y rehacer las comparaciones a voluntad y desde cualquier punto de su interfaz, de forma que pueda alternar de un modo a otro sin pérdida de tiempo y con total comodidad.

Todas las tablas y gráficas de Paramtext deberán ser modificadas en mayor o menor medida, desde la inclusión de simples columnas de totales hasta la combinación de datos parciales.

## *8.2 Formato de las tablas y gráficas*

Será necesario crear varios tipos de tablas combinadas dado que no todas requieren ser cambiadas de la misma manera. Algunas necesitarán incluir nueva información sin alterar el número de entradas (filas) mientras que otras aumentarán el número de filas y/o columnas.

Por un lado las gráficas deberán incluir la información de ambas obras pero por el otro deben seguir siendo fáciles de consultar por lo que no es una opción incluir gráficas con datos provenientes de demasiadas fuentes a la vez. Al existir ya gráficas con dos fuentes de datos (con palabras vacías y sin palabras vacías) deberían ampliarse hasta permitir cuatro pero eso la complicaría demasiado. Por eso se ha optado por incluir algún sistema que permita pasar de una gráfica (con palabras vacías) a otra (sin palabras vacías) donde en cada una de ellas se comparen ambas obras. El paso de una a otra deberá ser rápido y fluido.

## *8.3 Requerimientos funcionales – casos de uso*

### 8.3.1 Casos de uso

Un caso de uso es una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento que inicia un actor principal sobre el propio sistema. Se describen así las formas en las que el sistema puede ser utilizado sin necesidad de diseño o implementación, permitiendo detectar requerimientos funcionales en etapas tempranas del desarrollo.

Debe tenerse en cuenta que al ser una ampliación de un sistema ya definido y en funcionamiento tan sólo serán explicados aquellos que sean de nueva incorporación, los que hayan sido modificados o aquellos necesarios para el entendimiento de los anteriormente mencionados.

Los distintos pasos a seguir han sido:

- Identificación de los actores: se determinan los actores del sistema, tanto los humanos como los que no, que interactúan con el sistema.
- Establecimiento de objetivos por actor: se indican los objetivos que cada actor espera del sistema
- Relación de casos de uso: listado general de los casos de uso para cada actor
- Representación de casos de uso: muestra los distintos diagramas de casos de uso
- Descripción de casos de uso: descripción individual de los casos de uso que se han identificado

**Actores**

Al igual que en el programa original, Paramtext sigue tratando de igual manera a todos los actores externos de la aplicación, por lo que seguirá habiendo sólo uno; usuario. Sin embargo presenta herramientas nuevas para los administradores de la misma con lo que surge un nuevo actor secundario: administrador.

**Objetivos de los actores**

El objetivo del actor usuario consistirá en obtener información adicional del vocabulario y poder realizar comparaciones entre su texto y otros almacenados en nuestro servidor de forma rápida y sencilla.

#### 8.3.1.1 Lista de casos de uso

Además de los casos de uso del programa original, que pueden consultarse en su memoria (página 66 del pdf; página 54 del documento) el listado es como sigue.

| Actor principal | Caso de uso | Nº CU |
|---|---|---|
| Usuario | Seleccionar colección | CU 18 |
| | Seleccionar autor | CU 19 |
| | Seleccionar obra | CU 20 |
| | Iniciar comparación | CU 21 |
| | Ver información de la obra | CU22 |
| | Deshacer comparación | CU 23 |
| | Rehacer comparación | CU 24 |
| | Ver gráfica | CU 10* |
| | Gráfica sin palabras vacías | CU 25 |
| | Gráfica con palabras vacías | CU 26 |
| Administrador | Añadir colección | CU 27 |
| | Modificar colección | CU 28 |
| | Eliminar colección | CU 29 |
| | Añadir autor | CU 30 |
| | Modificar autor | CU 31 |
| | Eliminar autor | CU 32 |
| | Añadir información de obra | CU 33 |
| | Modificar información de obra | CU 34 |
| | Eliminar información de obra | CU 35 |

*Tabla 8.5*

#### 8.3.1.2 Diagramas de casos de uso:

Los casos de uso “Paramtext 1.0” representan aquellos que ya existían con anterioridad, salvo “ver gráfica” que ha tenido que ser incluido para enlazar con un nuevo caso de uso.

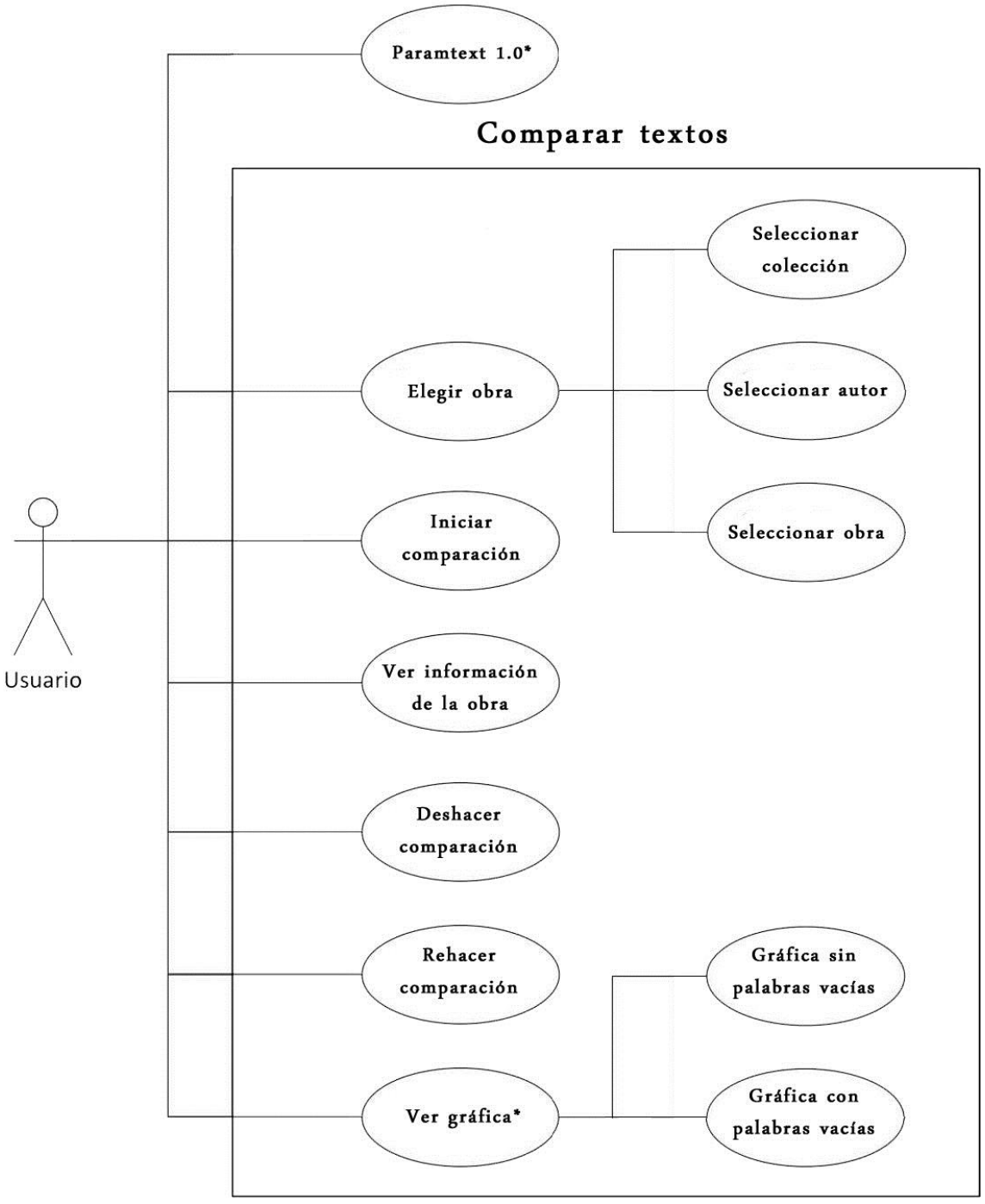



*Figura 8.1*

* No pertenecen a la ampliación de Paramtext.

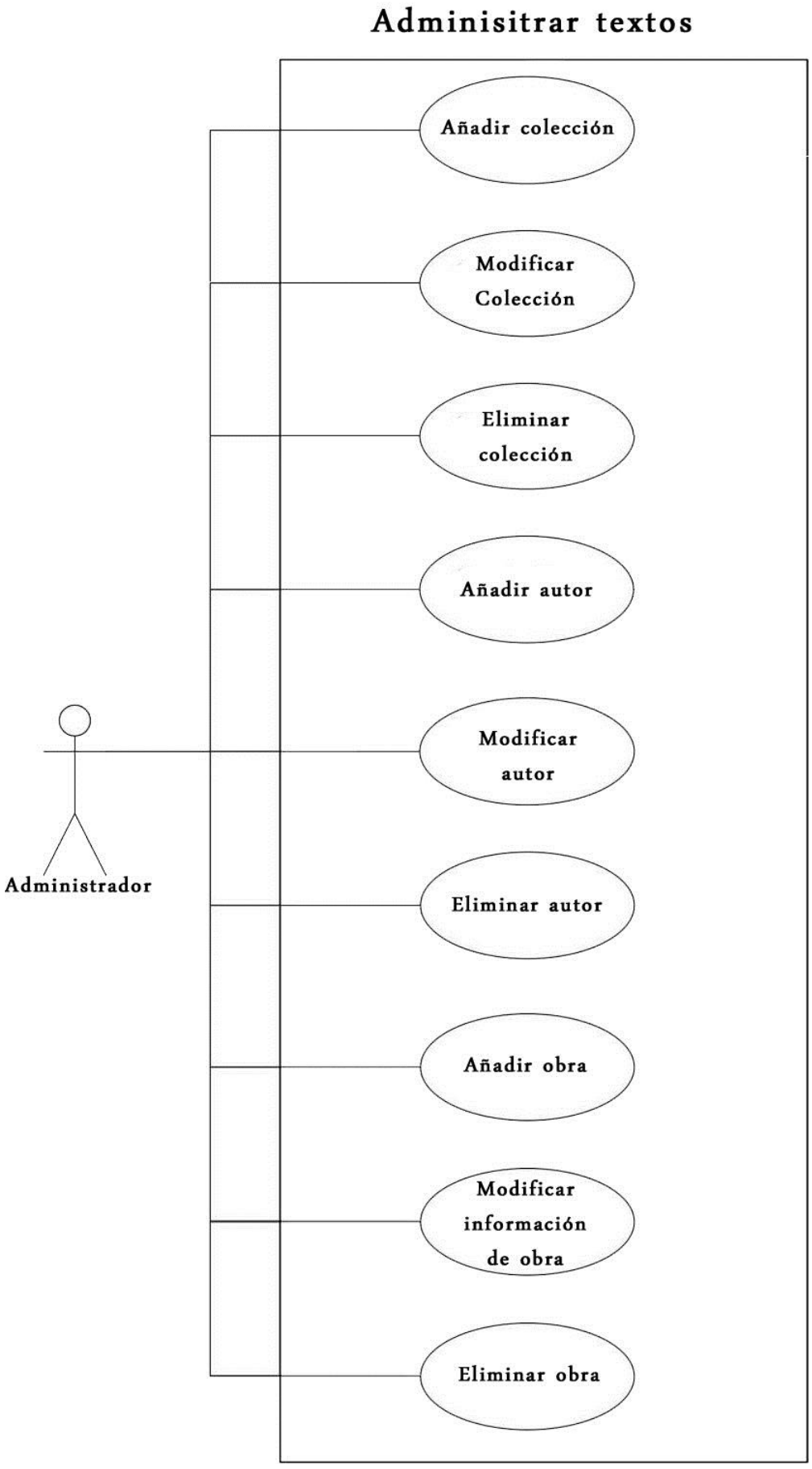

Adminisitrar textos
Añadir colección
Modificar Colección
Eliminar colección
Añadir autor
Modificar autor
Eliminar autor
Añadir obra
Modificar información de obra
Eliminar obra
Administrador

*Figura 8.2*

## Seleccionar colección CU18

Permite al usuario elegir entre las distintas colecciones que haya disponible en el archivo enlazado

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Situado en la página de comparación

**Parámetros**

- Lista de colecciones

**Flujo de ejecución**

1. Hacer clic en una de las colecciones que se muestran

**Excepciones**

**Postcondiciones**

- Estado: Se muestran los distintos autores disponibles para la colección que se haya seleccionado
- Estado: Se eliminan los anteriores autores del desplegable y se ocultan las obras e información de obra si se habían seleccionado con anterioridad

## Seleccionar autor CU19

Permite al usuario elegir entre los distintos autores que tengan alguna de sus obras dentro de la colección seleccionada o mostrar todas las obras disponibles independientemente de los autores

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Situado en la página de comparación
- Estado: Seleccionada una colección

**Parámetros**

- La colección seleccionada

**Flujo de ejecución**

1. Hacer clic en uno de los autores que se muestran o en la opción especial que los muestra todos

**Excepciones**

**Postcondiciones**

- Estado: Se muestran las distintas obras disponibles según el autor o todas si se ha seleccionado “todos” (si se ha seleccionado “todos” aparecerá el nombre del autor entre paréntesis junto al nombre de la obra).
- Estado: Se ocultan las anteriores obras e información de obra que se hubieran seleccionado con anterioridad

## Seleccionar obra CU20

Permite al usuario elegir entre las distintas obras que cumplen los filtros de colecciones y autores disponibles

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Situado en la página de comparación
- Estado: Seleccionada una colección
- Estado: Seleccionado un autor o “todos”

**Parámetros**

- El autor seleccionado o la opción “todos”

**Flujo de ejecución**

1. Hacer clic en una de las obras que se muestran

**Excepciones**

**Postcondiciones**

- Estado: Se muestra la información de la obra seleccionada y el botón “Comparar” que permite iniciar la comparación
- Estado: Se oculta la información de la obra que hubiera sido seleccionada con anterioridad

## Iniciar comparación CU21

Permite al usuario realizar la comparación entre su texto y el texto que ha seleccionado de entre los que ofrece la herramienta

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Situado en la página de comparación
- Estado: Seleccionada una colección
- Estado: Seleccionado un autor o “todos”
- Estado: Seleccionada una obra

**Parámetros**

- La obra seleccionada

**Flujo de ejecución**

1. Hacer clic en el botón “comparar”

**Excepciones**

**Postcondiciones**

- Estado: Entra en el modo comparación, se muestra la página “Métrica – Informe” del modo comparación.

## Ver información de la obra CU22

Permite al usuario ver toda la información del texto que ha seleccionado para comparar con el suyo

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Estar en modo comparación

**Parámetros**

- La información de la obra seleccionada

**Flujo de ejecución**

1. Situar el cursor sobre el icono en forma de pergamino junto al nombre del texto con el que se está comparando el suyo

**Excepciones**

**Postcondiciones**

- Estado: Ventana emergente que muestra la información de la obra

## Deshacer comparación CU23

Permite al usuario deshacer la comparación entre su texto y el texto que haya seleccionado

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Estar en modo comparación

**Parámetros**

**Flujo de ejecución**

1. Hacer clic en el icono en forma de equis junto al nombre del texto con el que se está comparando el suyo

**Excepciones**

**Postcondiciones**

- Estado: Sale del modo comparación y se sitúa en la misma página que se estaba viendo antes de deshacer la comparación

## Rehacer comparación CU24

Permite al usuario volver al modo comparación con la última obra con la que hubiera hecho una comparación

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Haber realizado al menos una comparación
- Estado: No estar en modo comparación

**Parámetros**

- La última obra seleccionada

**Flujo de ejecución**

2. Hacer clic en el icono de “Rehacer comparación”.

**Excepciones**

**Postcondiciones**

- Estado: Vuelve al modo de comparación con la última obra con la que se hubiera comparado situándose en la misma página en la que se estaba cuando se apretó el icono de rehacer la comparación.

# Ver gráfica CU10

Permite al usuario ver el resultado escogido en formato de gráfica

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Página de resultados cargada

**Parámetros**

**Flujo de ejecución**

1. En la página de resultados el usuario selecciona la opción de gráfica

**Excepciones**

**Postcondiciones**

- Estado: Página de resultados cargada
- Estado: Resultados en formato de gráfica

**Notas**

- Existen resultados como la media, moda, mediana, etc. que al ser de carácter unitario no tendrán disponibles su visualización en modo gráfica

## Grafica sin palabras vacías CU25

Permite al usuario ver las gráficas sin tener en cuenta las palabras vacías que haya en el texto

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Estar en modo comparación

**Parámetros**

**Flujo de ejecución**

1. Activar la casilla “sin palabras vacías” haciendo clic en ella

**Excepciones**

**Postcondiciones**

- Estado: Página de resultados cargada
- Estado: Se muestra la gráfica sin tener en cuenta las palabras vacías que haya en el texto

## Grafica con palabras vacías CU26

Permite al usuario volver al modo comparación con la última obra con la que hubiera realizado la comparación

**Precondiciones**

- Estado: Archivo parametrizado
- Estado: Estar en modo comparación

**Parámetros**

**Flujo de ejecución**

1. Desactivar la casilla “sin palabras vacías” haciendo clic en ella

**Excepciones**

**Postcondiciones**

- Estado: Página de resultados cargada
- Estado: La gráfica muestra las palabras vacías

## Añadir colección CU27

Permite al administrador incluir una nueva colección entre las que el usuario pueda elegir en la página de iniciar comparación

**Precondiciones**

**Parámetros**

**Flujo de ejecución**

1. Escribir la nueva colección y XML asociado en el archivo de texto lista_colecciones.txt

**Excepciones**

**Postcondiciones**

- Estado: Se muestra una nueva colección al usuario en la página de iniciar comparación que incluye sus propias obras

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con recargar la página.

# Modificar colección CU28

Permite al administrador modificar el nombre o el archivo XML asociado a las colecciones que se muestran al usuario

**Precondiciones**

- Colección creada

**Parámetros**

**Flujo de ejecución**

1. Escribir la nueva colección y XML asociado en el archivo de texto lista_colecciones.txt

**Excepciones**

**Postcondiciones**

- Estado: Se actualiza la información de la colección modificada

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con recargar la página.

## Eliminar colección CU29

Permite al administrador eliminar una de las colecciones que se muestran al usuario en la página de iniciar comparación

**Precondiciones**

- Colección creada

**Parámetros**

**Flujo de ejecución**

1. Borrar la colección y el XML asociado que se encuentra en el archivo de texto lista_colecciones.txt

**Excepciones**

**Postcondiciones**

- Estado: Deja de mostrarse al usuario la colección eliminada

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con recargar la página.

## Añadir autor CU30

Permite al administrador añadir nuevos autores

**Precondiciones**

**Parámetros**

**Flujo de ejecución**

1. Crear carpeta con el nombre del autor normalizado dentro de la carpeta de estadísticas
2. Incluir en el archivo XML de la colección alguna obra del autor

**Excepciones**

**Postcondiciones**

- Estado: Aparece un nuevo autor al seleccionar las colecciones que incluyan alguna de sus obras

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con volver a seleccionar la colección o recargar la página.

## Modificar autor CU31

Permite al administrador modificar un autor existente

**Precondiciones**

- Autor creado

**Parámetros**

**Flujo de ejecución**

1. Modificar el nombre de la carpeta
2. Modificar el nombre del autor en los archivos XML donde aparezca

**Excepciones**

**Postcondiciones**

- Estado: Modifica el autor

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con volver a seleccionar la colección o recargar la página.

## Eliminar autor CU32

Permite al administrador eliminar un autor existente de una colección o eliminarlo por completo de la lista de autores

**Precondiciones**

- Autor creado

**Parámetros**

**Flujo de ejecución**

1. Eliminar las obras del autor del XML asociado a una o varias colecciones
2. (Opcional: para su completa eliminación) Eliminar la carpeta del autor con las estadísticas de sus obras

**Excepciones**

**Postcondiciones**

- Estado: Elimina el autor de una colección o completamente

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con volver a seleccionar la colección o recargar la página.

## Añadir obra CU33

Permite al administrador añadir nuevas obras

**Precondiciones**

- Estado: Debe existir una carpeta de autor

**Parámetros**

- Nombre de la carpeta de autor

**Flujo de ejecución**

1. Crear archivo de estadísticas usando Paramtext
2. Copiar carpeta de estadísticas dentro de la carpeta del autor con el nombre de la obra normalizada
3. Incluir información de la obra en los archivos XML de las colecciones

**Excepciones**

**Postcondiciones**

- Estado: Aparece la obra indicada en la/s colecciones donde se haya incluido

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con volver a seleccionar el autor o recargar la página.

## Modificar información de la obra CU34

Permite al administrador modificar la información de las obras existentes de forma independiente

**Precondiciones**

- Obra creada

**Parámetros**

**Flujo de ejecución**

1. Modificar información de la obra en los archivos XML de las colecciones

**Excepciones**

**Postcondiciones**

- Estado: Cambia la información de la obra

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con volver a seleccionar la obra o recargar la página.

## Eliminar obra CU35

Permite al administrador eliminar una obra existente de una colección o eliminarla completamente del sistema

**Precondiciones**

- Obra creada

**Parámetros**

**Flujo de ejecución**

1. Elimina la información de la obra en el archivo XML de la colección
2. (Opcional: para su completa eliminación) Eliminar la carpeta de estadísticas de la obra

**Excepciones**

**Postcondiciones**

- Estado: Elimina la obra de una colección o completamente

**Nota**

- No es necesario detener el programa para que tenga efecto la actualización, basta con volver a seleccionar la obra o recargar la página.

## *8.4 Requerimientos no funcionales*

A continuación se listan y detallan los requerimientos no funcionales que debe cumplir el sistema y que permiten comprender las propiedades del software a desarrollar.

Además de los requerimientos no funcionales incluidos en el programa original que se pueden ver en su propia memoria (página 87 del PDF; página 75 del documento) se han incluido los siguientes:

- **Facilidad de ampliación**

  El nuevo software debe permitir que se incluyan, modifiquen y eliminen nuevas obras, colecciones y autores sin necesidad de entrar en el código del programa.

- **Homogeneidad**

  El programa debe ampliarse manteniendo en la medida de lo posible la estructura y el metodología utilizada para evitar generar un código totalmente independiente que no tenga nada que ver con las bases sentadas en el original

- **Continuidad**

  Deben poder realizarse las ampliaciones y modificaciones en las colecciones y las obras sin necesidad de detener la herramienta, de forma que los usuarios puedan seguir disfrutando de ella mientras se producen.

# 9 Diseño

Una vez se han sentado las bases de lo que se pretende conseguir mediante el análisis es posible empezar con la fase de desarrollo. La primera tarea en esa dirección es el diseño y su importancia es tal que un error en esta etapa puede dar lugar a la pérdida de semanas o meses de trabajo. En ella es necesario definir la estructura y comportamiento que debe tener el sistema de manera que se cubran los requerimientos de la fase anterior.

## *9.1 Arquitectura cliente-servidor*

Dado que la ampliación del software se presentará al usuario final mediante una modificación de la página web del sistema original se mantendrá la arquitectura existente. Dicha arquitectura puede consultarse en la memoria original del sistema (página 89 del PDF; página 77 del documento).

## *9.2 Diseño del sistema*

En el análisis se ha determinado que la ampliación del sistema debe adaptarse al software existente, de tal forma que a menos que se pretenda, no se distinga uno del otro. La primera tarea consisten en eliminar parte del código y utilizar un servicio en su lugar, posteriormente habrá que añadir la información extraída de la RAE y crear un nuevo algoritmo capaz de extraer los distintos n-gramas del texto. Por último habrá que añadir la herramienta que permita comparar los textos a lo largo de todo el programa.

El sistema inicial está dividido en dos subsistemas, una aplicación de consola para la extracción de información (PMT) y otra aplicación de tipo web para las funciones de estructuración e interfaz para el usuario (Paramtext). Dado que se pretende tanto el tratamiento y reestructuración de nueva información como la adaptación de la interfaz de usuario será necesario realizar modificaciones en ambos

### *9.2.1 Modificaciones y ampliaciones al subsistema PMT*

El subsistema actual es el encargado de realizar realmente el trabajo duro del sistema. Consiste en una aplicación de consola (ejecutable) con la finalidad de realizar todas las tareas relacionadas con la Parametrización morfológica de textos.

Para en buen entendimiento del trabajo realizado se ha tenido que incluir parte de las explicaciones originales del sistema. Si se desea puede consultarse la estructura completa del PMT en la memoria original (página 91 del PDF; página 79 del documento).

#### 9.2.1.1 Silabeador

Este módulo se encarga de realizar la separación de las palabras en sílabas para ser mostradas en el apartado de vocabulario. Al ser un código incrustado, cualquier modificación del código debía ser realizada dentro del propio programa y en todos aquellos lugares externos al proyecto donde se quisiera actualizar la herramienta.

Por tanto la modificación consistirá por un lado en el uso de un servicio externo al proyecto que devuelve cierta información, y por otro lado la utilización de dicha información para separar las sílabas de la palabra mediante un algoritmo propio.

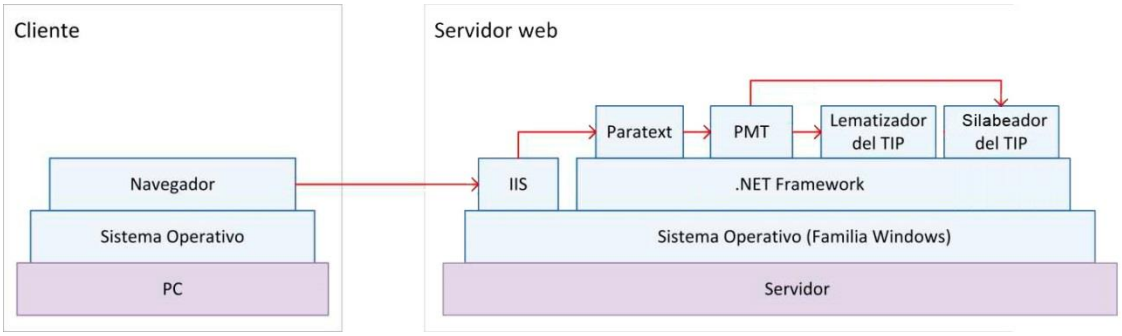



*Figura 9.1: Diseño arquitectónico del sistema.*

Dado que el silabeador del TIP es un programa externo al propio proyecto no se entrará en detalle acerca de su funcionamiento.

### 9.2.1. 2 Nueva información Léxica

Al disponer de nuevos datos será necesario ampliar las estructuras que los almacenaban y el algoritmo que realizaba las peticiones. En primer lugar se seguirán extrayendo del servicio lematizador (tal y como se hacía anteriormente) aunque ahora la cantidad de información seá mayor por lo que se incorporarán las nuevas funciones de adquisición de información en la clase *Analizador Morfológico* y los campos necesarios para almacenarlos en la clase *Información Morfológica*.

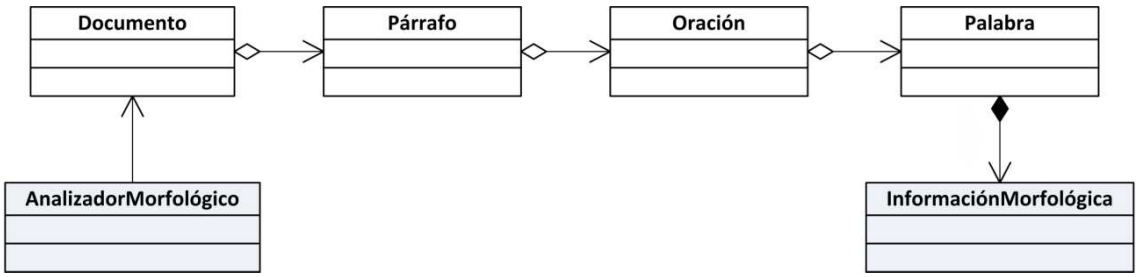



*Figura 9.2: Diagrama de clases del módulo Lematización.*

Adicionalmente será necesario serializar también la nueva información para que pase al archivo XML que hace de enlace entre los subsistemas PMT y Paramtext por lo que se deberá modificar también dichas estructuras para que la serialización se produzca también en ellas. De esta forma podremos usar los mismos archivos de serialización en lugar de tener que crear otros nuevos.

El proceso de serialización necesita de un proceso inverso a su llegada al subsistema Paramtext por lo que deberá tenerse en cuenta para su correcto funcionamiento

### 9.2.1.3 Nueva información n-gramas

Para intentar mantener la arquitectura y estructuras pertenecientes al programa original se ha optado por usar la misma estructura de “Documento - Párrafos – oraciones - palabras” existente, modificándola a “Documento - Párrafos – oraciones – n-gramas - palabras”.

Para almacenar los n-gramas se utilizarán las mismas estructuras que para las oraciones aunque será necesario modificarlas y crear varias de ellas para que acepten distintos tamaños (n-gramas de grado entre dos y cinco) y que los mantengan separados para poder mostrarlos según su tamaño.

Por último tendrá que crearse un proceso de serialización adicional para tratar las nuevas estructuras y un nuevo fichero XML donde almacenarlos dado que su tamaño y su naturaleza no encajan con los existentes hasta el momento.

### *9.2.2 Modificaciones y Ampliaciones al subsistema Paramtext*

El subsistema Paramtext es una aplicación tipo web a la que acceden los usuarios para solicitar la Parametrización morfológica de sus texto. La modificación de la interfaz será considerable aunque se mantendrá el estilo general de la misma. Se mantendrá el Modelo Vista Controlador pues no es necesario cambiarlo para realizar la ampliación.

Deberá ser modificada para incluir la nueva información extraída de la RAE y poder mostrar los n-gramas que componen el texto. Sin embargo la mayoría de los cambios afectarán al nuevo modo de comparación en el cual todas las tablas y gráficas serán modificadas en mayor o menor medida para poder combinar los datos provenientes del texto del usuario y de aquellos que están almacenados en nuestros servidores.

Por último se incluirá una herramienta de selección de obra con la que comparar el texto que permite discriminar entre distintas colecciones y autores con facilidad.

#### 9.2.2.1 Nueva información Léxica

La información extraída de la Rae mediante las herramientas auxiliares definidas en el capítulo siete de esta memoria tendrán su propio apartado dentro de la información de vocabulario que se ofrece. Por tanto será necesario realizar los siguientes ajustes.

1. Un nuevo desplegable al situarse sobre “Vocabulario”
2. Creación de una nueva tabla dinámica para introducir la información
3. Adaptación de los datos recibidos para que sean de fácil comprensión para el usuario
    a. Cambiar del código numérico de la categoría gramatical por una abreviación de la propia categoría.
    b. Desplegable con su nombre completo para las abreviaciones de la categoría gramatical
    c. Cambiar el 0 y el 1 por “No” o “Sí” en las columnas “Antigua” y “Aparece en la RAE”

1) Actualmente se accede al vocabulario mediante un enlace directo en la parte superior de la aplicación. Para poder dividir la información en varios apartados se debe añadir un menú desplegable entre los que poder elegir si mostrar la información estadística (la que existía anteriormente) o la información léxica.


*Figura 9.3: Menú desplegable del vocabulario*

2) La forma de mostrar la información será mediante una nueva tabla dinámica. Dicha tabla permitirá reorganizar el orden de las filas haciendo clic en el encabezado de una columna. Los datos que se mostrarán serán los siguientes:

1. *Forma canónica:* Indica su forma canónica
2. *Categoría gramatical:* Muestra una abreviatura de la categoría gramatical
3. *Etimologías:* Presenta el número de etimologías
4. *Total de acepciones:* Señala el número total de acepciones
5. *Posición de la acepción:* Indica la posición que tiene la acepción en el diccionario de la RAE
6. *Número de entradas en la acepción:* Muestra cuantas entradas tiene esa acepción en el diccionario de la RAE
7. *Antigua:* Señala si es considerada como antigua por la RAE
8. *Aparece en la RAE:* Informa si aparece en el diccionario de la RAE

3a) El sistema trabaja con los códigos de las categorías gramaticales por tanto es necesario realizar una conversión de dicho valor numérico a la categoría que representa.

3b) Debido a la gran longitud de muchas de las categorías gramaticales se ha optado por usar la abreviación de las mismas a la hora de mostrarlas en la tabla. Así mismo, para facilitar su comprensión se proporcionará la posibilidad de ver el nombre completo en un desplegable si se sitúa el cursor sobre la abreviación.

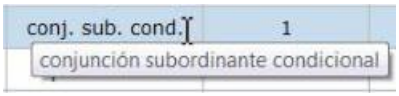



*Figura 9.4: tooltip de la categoría gramatical sin abreviar*

3c) El sistema almacena internamente los valores de cero como si fueran un “no”, y los valores de uno como si fueran un “sí”. Para no mostrar ceros y unos al usuario se debe realizar el cambio correspondiente.

### 9.2.2.2 Nueva información N-gramas

Al existir un nuevo fichero XML que tendrá su propio espacio en el sistema de carpetas del programa deberá modificarse la clase modelo de enrutamiento de modo que pueda albergar los nuevos archivos y carpeta.

Al igual que la nueva información léxica, los n-gramas tendrán su propio apartado en la sección de vocabulario. El total de los cambios comprende lo siguiente:

1. Una nueva entrada en el desplegable “Vocabulario” para acceder a los n-gramas
2. Creación de cuatro tablas dinámicas para introducir la información
3. Organizarlas mediante un sistema de pestañas para acceder rápidamente de una tabla a otra

1) Se pretende ampliar el acceso a la información del vocabulario mediante un menú desplegable que separará las estadísticas de la información léxica, por lo que se pretende usar el mismo sistema para incluir una nueva línea de acceso para los n-gramas.

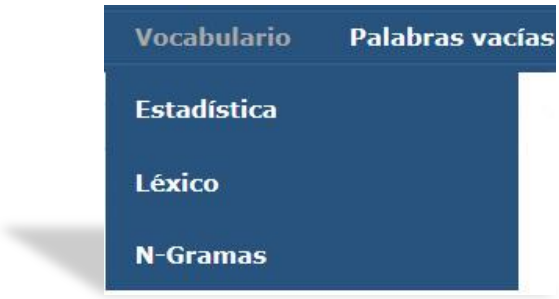



*Figura 9.5: Menú desplegable ampliado del vocabulario*

2) La forma de mostrar la información será mediante cuatro nuevas tabla dinámica. Dichas tablas permitirán reorganizar el orden de las filas haciendo clic en el encabezado de cada columna. Los datos que se mostrarán serán los siguientes:

1. *N-grama*: Presenta el n-grama
2. *Frecuencia (texto):* Muestra el número de veces que aparece en el texto
3. *Primera aparición:* Indica la primera aparición del n-grama
4. *Centro de gravedad:* Define el centro de gravedad del n-grama en el texto, lo que ayuda a saber hacia que parte del texto puede encontrarse con mayor frecuencia.

3) Dado que van a mostrarse n-gramas de distintos tamaños, entre dos y cinco, es una buena idea el poder acceder a ellos de forma rápida y cómoda de modo que se implantará un sistema de pestañas para pasar de uno a otro.


*Figura 9.6: Pestañas del apartado N-gramas*

### 9.2.2.3 Modificación de la interfaz en el modo original (no el de comparación)

Aunque la mayor parte de la interfaz quedará tal y como está se han realizado algunos cambios menores.

1. Se añadirán nuevos estilos en el formato de las tablas para mejorar la visualización de los mismo
2. Se incluirá una nueva línea de información tras retornar del modo comparación que permitirá volver a él con un simple clic

1) Algunas tablas muestran la información de las filas a dos niveles porque alguno de los campos es demasiado pequeño para la cantidad de información que sostienen mientras que otros tienen espacio de sobra. Otras tienen algunas columnas demasiado juntas entre ellas mientras que las demás están demasiado separadas. Se añadirán estilos para ajustar las tablas según sea necesario.

| Flexión verbal | Frecuencia |
|---|---|
| Infinitivo | 4 |
| Participio | 1 |
| 1ª persona del singular del presente de indicativo | 1 |
| 3ª persona del singular del presente de indicativo<br>2ª persona del singular del presente de indicativo (usted) | 4 |

*Figura 9.7: Tabla con filas a dos niveles*

| Flexión verbal | Frecuencia |
|---|---|
| Infinitivo | 4 |
| Participio | 1 |
| 1ª persona del singular del presente de indicativo | 1 |
| 3ª persona del singular del presente de indicativo<br>2ª persona del singular del presente de indicativo (usted) | 4 |

*Figura 9.8: Tabla con el nuevo estilo para evitar los dos niveles*

2) Para permitir un rápido traspaso del modo original al modo comparación en Paramtext se optará por incluir una línea de información nueva que incluya información sobre la última obra comparada y un icono sobre el que hacer clic que permitirá volver a la última comparación que se haya realizado.

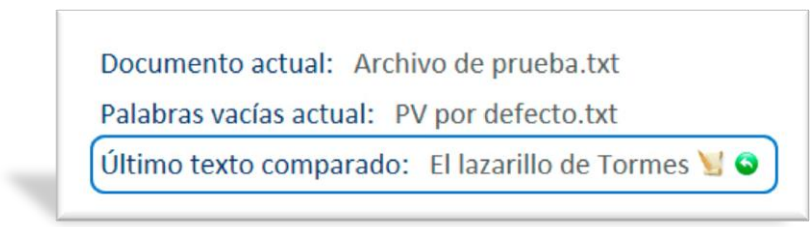



*Figura 9.9: Nueva línea de información y enlace de retorno al modo de comparación*

#### 9.2.2.4 ParamTextComp

Este nuevo modo de consulta requiere la modificación de todas las gráficas y tablas del sistema para que presente la información del usuario y de la obra seleccionada para comparar simultáneamente.

Se necesita un nuevo apartado en donde el usuario pueda elegir entre las obras existentes, que estarán divididas en colecciones y estas a su vez por autores (aunque el usuario siempre podrá elegir ver todas las obras a la misma vez independientemente de los autores). Este sistema deberá permitir al administrador realizar modificaciones sin necesidad de entrar en el código del programa.

Por último el usuario deberá poder consultar la información de la obra en todo momento sin necesidad de cambiar de apartado y por supuesto deshacer la comparación en cualquier instante con un simple clic.

**Formatos de las nuevas tablas y gráficas:**

La cantidad y variedad en los formatos y naturaleza tanto de las tablas como de las gráficas utilizadas por Paramtext ha dado lugar a varios acercamientos distintos a la hora de modificar las tablas para realizar las comparaciones. A continuación se muestran los distintos tipos de ampliaciones y combinaciones así como ejemplo simplificados de muestra.

1. Añadir columnas totales: este tipo de tabla será la más sencilla porque sólo habrá que añadir una nueva columna de totales y rellenarla con los datos de la obra que el usuario quiera comparar.

2. Añadir columnas intercaladas de datos y porcentajes: Al encontrarse intercaladas y no tener todas las tablas el mismo tamaño ni provenir los datos de un mismo origen no puede generarse un algoritmo genérico. Por tanto debe hacerse la inclusión de cada columna y fila una a una.

3. Combinar datos en base a uno de ellos: al ser distintas gráficas y parámetros los que se deben tener en cuenta para realizar la combinación y ser variable el tamaño de las tablas y por la alta variedad de casos y la posibilidad de encontrarnos valores en una obra que no están en la otra que deben combinarse, este será el formato más complicado. En el ejemplo se puede ver una tabla combinada según el número de caracteres de las palabras de las distintas obras, nótese que en una no hay palabras de nueve y trece caracteres mientras que en la otra no las hay de diez.

4. Combinación de datos y metadatos: El cuarto formato de tabla se utiliza para dar información acerca de los datos que se muestran y no para comparar los datos entre sí. En el ejemplo, las formas canónicas tienen los mismos valores tanto si pertenecen a una obra como a la otra, por lo que duplicar la información no tiene sentido, en su lugar se incluye una nueva columna que indica en qué obra se encuentra dicha forma o si está en ambas.

Las gráficas son un elemento importante a la hora de interpretar datos si necesidad de saber los valores exactos, mediante un sistema de "tamaños" se determina fácil e intuitivamente aquello que sobresale o se queda corto. Sin embargo son un arma de doble filo pues en el momento en el que la cantidad datos procedentes de distintas fuentes va aumentado la facilidad para interpretar lo que se muestra disminuye radicalmente hasta el punto de no verse más que un conglomerado de barras y líneas que no dicen nada. Es por eso que las gráficas serán modificadas de una forma u otra dependiendo de la cantidad de fuentes distintas que originalmente le aportaban valores. Dos han sido los tipos de modificaciones que han sufrido dependiendo de si originalmente disponían de una entrada de datos o si procedían de dos fuentes distintas:

1. De una a dos: si la gráfica tenía un único tipo de valor, este se mostrará junto con el de la obra con el que se quiere comparar. Su visualización conjunta ayuda a establecer la diferencia cuantitativa de un simple vistazo sin que la consulta de la misma se complique.

2. De dos a cuatro: si la gráfica tenía ya una comparación anterior, como la que se produce al comparar los valores cuando se tienen en cuenta las palabras vacías y cuando no, entonces deberán generarse dos gráficas distintas que puedan intercambiarse al marcar la casilla correspondiente. Una de las gráficas mostraría la comparación entre ambas obras cuando se tienen en cuenta las palabras vacías y la otra cuando no.

El listado detallado de cambios requeridos en tablas y gráficas es el siguiente:

**Métrica: informe**

- *Totales con palabras vacías:* añadir columna de totales
- *Totales sin palabras vacías*: añadir columna de totales
- *Promedios con palabras vacías:* añadir columnas intercaladas de datos y porcentajes
- *Promedios sin palabras vacías:* añadir columnas intercaladas de datos y porcentajes

**Métrica: gráficas**

- *Palabras de N caracteres:* combinación de datos según el *nº de caracteres* y generación de nuevas tablas; modificación de las gráficas de dos a cuatro.
- *Oraciones de N caracteres*: combinación de datos según el *nº de caracteres* y generación de nuevas tablas; modificación de las gráficas de dos a cuatro.
- *Párrafos de N caracteres:* combinación de datos según el *nº de caracteres* y generación de nuevas tablas; modificación de las gráficas de dos a cuatro.
- *Oraciones de N caracteres:* combinación de datos según el *nº de palabras* y generación de nuevas tablas; modificación de las gráficas de dos a cuatro.
- *Párrafos de N caracteres:* combinación de datos según el *nº de palabras* y generación de nuevas tablas; modificación de las gráficas de dos a cuatro.
- *Párrafos de N caracteres:* combinación de datos según el *nº de oraciones* y generación de nuevas tablas; modificación de las gráficas de dos a cuatro.
- *Distribución por frecuencia:* combinación de datos según la *frecuencia de las palabras* y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Distribución por centro de gravedad:* combinación de datos según el *centro de gravedad de las palabras* y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Distribución por primera aparición:* combinación de datos según la *primera aparición de las palabras* y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Distribución en el corpus:* combinación de datos según la *frecuencia en el corpus* y generación de nuevas tablas; modificación de las gráficas de una a dos.

**Morfología: informe**

- *Totales por categoría gramatical:* añadir columna de totales
- *Promedios por categoría gramatical:* añadir columnas intercaladas de datos y porcentajes
- *Totales por flexión verbal:* añadir columna de totales
- *Totales por flexión no verbal:* añadir columna de totales

**Morfología: gráficas**

- *Categorías gramaticales:* combinación de datos según la categoría gramatical y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Flexiones verbales:* combinación de datos según las distintas flexiones verbales y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Flexiones no verbales:* combinación de datos según las distintas flexiones no verbales y generación de nuevas tablas; modificación de las gráficas de una a dos.

- *Categorías gramaticales - verbos:* combinación de datos según el nº de verbos y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - sustantivos:* combinación de datos según el nº de sustantivos y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - adjetivos:* combinación de datos según el nº de adjetivos y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - adverbios:* combinación de datos según el nº de adverbios y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - pronombres:* combinación de datos según el nº de pronombres y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - preposiciones:* combinación de datos según el nº de preposiciones y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - artículos:* combinación de datos según el nº de artículos y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Categorías gramaticales - conjunciones:* combinación de datos según el nº de conjunciones y generación de nuevas tablas; modificación de las gráficas de una a dos.
- *Morfología - Palabras no reconocidas:* añadir columnas intercaladas de datos y porcentajes.

**Vocabulario: Estadísticas**

*Con y sin palabras vacías:* Combinación de datos según la palabra e intercalado de datos y porcentajes.

**Vocabulario Léxico**

*Léxico:* determinar en qué obras aparecen las formas canónicas de ambos textos.

**Vocabulario N-gramas**

*Tamaño 2, 3, 4 y 5:* Combinación de datos según los n-gramas de distintos tipos e intercalado de datos

### Apartado de selección de obras:

La selección de las obras deberá ser sencilla, rápida y versátil. El usuario debe encontrar lo que necesita con un par de clics y tener toda la información necesaria a la vista. Se creará una nueva página en la que mostrar la interfaz principal de la nueva herramienta y su distribución será la siguiente:

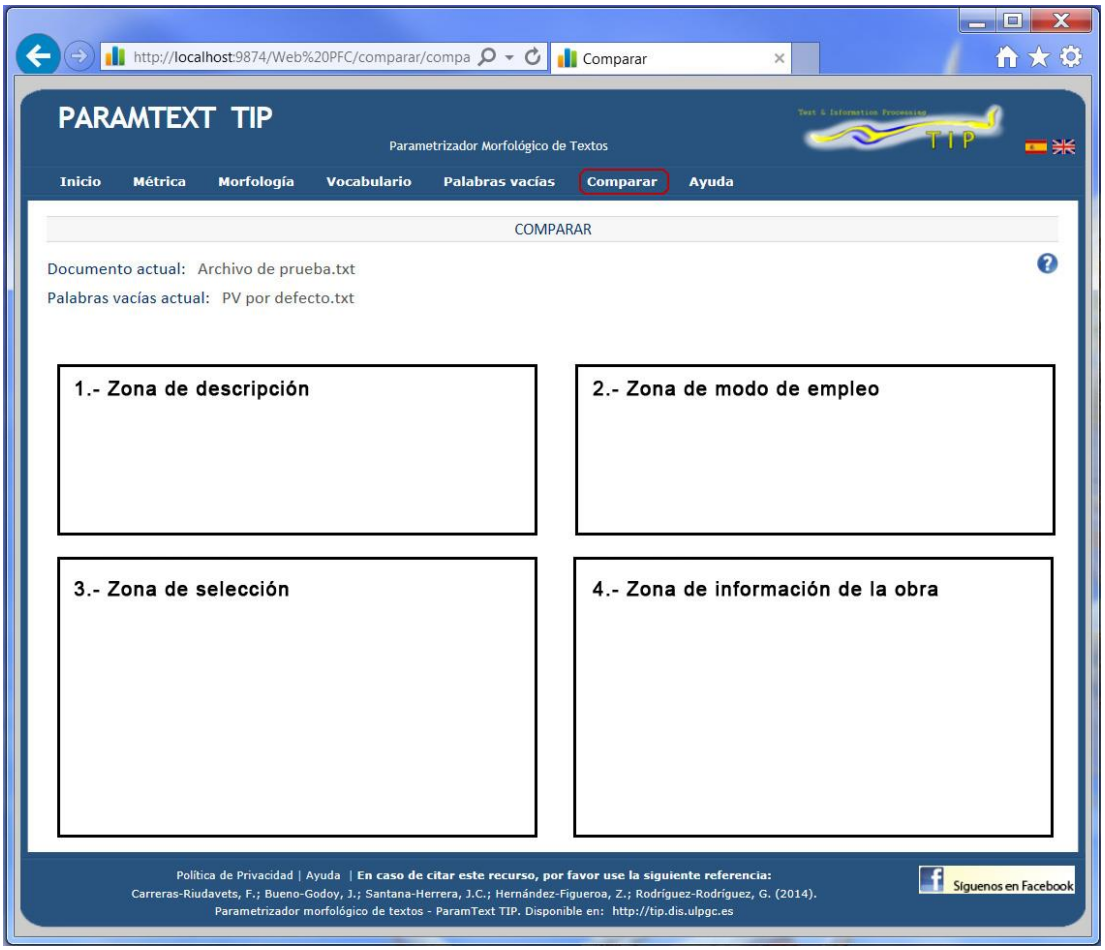



*Figura 9.17: prototipo de la página de acceso al modo de comparación*

#### Zona de descripción

Debe explicar a groso modo la utilidad de la herramienta, para que los usuarios sepan lo que pueden esperar de ella. Deberá ocupar como mucho un párrafo y no pretende ser una descripción exhaustiva.

#### Zona modo de empleo

Será un resumen de la metodología que se debe seguir para realizar la comparación. Su uso en detalle estará incluido en la ayuda de la herramienta, donde se marcará con la etiqueta V.2014 lo que significa que pertenece a la versión de ese año.

**Zona de selección**

En ella el usuario seleccionará la obra con la que quiere comparar su texto. Aunque inicialmente sólo será visible una de ellas, esta zona estará dividida en tres partes:

- *Colecciones disponibles*: Agrupará las obras según un elemento común, como puede ser que pertenezcan a la literatura clásica o que sean escritos periodísticos.
- *Autores*: Permitirá elegir entre los distintos autores que poseen obras en la colección seleccionada o seleccionar "todos" para que aparezca un listado completo de las obras de esa colección.
- *Obras*: Mostrará las obras que cumplen los filtros seleccionados. En caso de haber seleccionado la opción especial "todos" en la selección de autor el nombre de las obras irá seguido del nombre del autor entre paréntesis.

**Zona de información de la obra**

En esta zona es donde se mostrará la información de la obra seleccionada en la zona de selección. Podrá incluir tantos campos como sean necesarios y dependerá del archivo XML asociado a la colección a la que pertenece. En la parte inferior se situará el botón que da lugar a la comparación.

**Repositorio de obras**

El sistema de almacenamiento de las obras que pueden seleccionarse para comparar se basará en un sistema de carpetas en el interior del programa gestionado por archivos XML donde se almacenará la información que se mostrará al usuario cuando seleccione la obra así como otros datos relevantes para el buen funcionamiento de la herramienta de comparación como puede ser la abreviación del nombre del autor para mostrarse en las tablas de comparación.

Se añadirá un archivo "Readme.txt" que mostrará su correcto funcionamiento e incluirá las instrucciones para añadir, borrar o modificar cualquiera de los parámetros de la herramienta de selección de obras y los datos de las obras.

**Los pasos resumidos para incluir nuevos elementos serán los siguientes:**

1. Agregar una nueva colección en "App_Data/libros/colecciones/lista_colecciones.txt" [Nombre archivo XML que contiene las obras de la colección]:[Nombre que se quiere mostrar al usuario]

2. Crear archivo XML que contenga las obras de la colección en la misma carpeta

3. Crear una carpeta con el NOMBRE NORMALIZADO de cada autor incluido en el archivo XML dentro de "App_Data/libros/estadisticas"

4. Crear una carpeta con el NOMBRE NORMALIZADO de cada obra incluida en el archivo XML dentro de la carpeta del autor correspondiente "App_Data/libros/estadisticas/[Autor de la obra]/[Nombre de la obra]

5. Situar las estadísticas de cada obra dentro de su carpeta correspondiente

**Incluir nuevas colecciones**

La herramienta generará las colecciones de forma dinámica y será un número ilimitado de ellas.

Para incluir las nuevas colecciones bastará con añadir una nueva línea en el fichero "lista_colecciones.txt" situado en "App_Data/libros/colecciones" y su correspondiente fichero de colección. El formato del fichero TXT deberá ser exactamente el que se describe; nombre del fichero XML (con extensión incluida) seguida (sin espacios) de la descripción que se quiere mostrar al usuario (admite espacios y cualquier símbolo).

Ejemplo de fichero lista_colecciones.txt:

- literatura_clásica.xml:Literatura Clásica
- literatura_moderna.xml:Literatura Moderna

Ejemplo de modificación en dicho fichero:

- literatura_clásica.xml:Literatura Clásica
- literatura_moderna.xml:Literatura Moderna
- pergaminos_super_viejos.xml:Pergaminos Antiguos
- piedra_para_escribir.xml:Tablillas de Piedra
- de_la_A_a_la_Z.xml:Abecedario Extendido

Los nombres de los archivos XML NO necesitan coincidir en ningún caso con el alias que se le pretende dar a ver al usuario (aunque en recomendable cierto parecido para una fácil identificación por nuestra parte).

La ordenación tampoco es necesaria pues se mostrará al usuario siempre en estricto orden alfabético.

El algoritmo leerá aquellas líneas que incluyan la cadena "X.xml:" (sin comillas) donde "X" será una cadena de tamaño uno (1) o superior, e ignorará el resto del fichero.

**Nuevos campos en archivos XML:**

El archivo XML que lista las obras de una colección podrá ser editado para incluir nuevos atributos que se mostrarán al usuario cuando seleccione una obra. Para hacerlo bastará con incluir un nuevo campo dentro de la etiqueta <data...> teniendo en cuenta lo siguiente:

- Titulo siempre tiene que ser el primer campo de la etiqueta
- Los nombres que identifican los campos no admitirán espacios ni símbolos pero el que se muestra al usuario sí.

De tal manera que, por ejemplo, se podrá añadir el nombre de campo "Palabra", pero no "Número de Palabras", sin embargo si es posible que el campo que se muestre al usuario se llame "Número de Palabras".

Así mismo, no todos los libros (ni siquiera de la misma colección) necesitarán tener los mismos atributos dentro de la etiqueta <data...> (salvo Titulo y Autor que siempre debe estar). De esta manera se podrán personalizar los atributos que se quieren mostrar para cada libro o colección, pudiendo usar por ejemplo, unos para "Literatura Clásica" y otros totalmente distintos para "Artículos Periodísticos" o "Entradas de Blogs".

**Nuevos autores:**

Las carpetas para los autores se situarán en "App_Data/libros/estadisticas/" y siguen las mismas reglas de nomenclatura que las obras. Dichas reglas se detallan en el siguiente apartado.

**Nuevas obras:**

La herramienta generará las obras disponibles de forma dinámica, tan sólo habrá que incluir la descripción de la misma en el fichero de colección correspondiente manteniendo su formato (se recomienda "copiar y pegar" para realizar los añadidos).

Los datos de la obra deberán incluirse en su propia carpeta, situada dentro de la de su autor, que estará en "App_Data/libros/estadisticas/[Nombre del Autor]/".

La carpeta de obra deberá tener el mismo nombre que el título de la obra salvo por aquellos caracteres que pueden generar conflicto a la hora de compilar el proyecto. A groso modo, las carpetas solo podrán usar los siguientes caracteres:

- A-Z
- a-z
- 0-9
- punto (.)
- coma (,)

De tal forma que el programa generará automáticamente la siguiente normalización en los títulos del archivo XML:

- Las letras con tilde serán cambiadas automáticamente por letras sin tilde
- La "ñ" será cambiada por la "n"
- Cualquier otro símbolo será eliminado

Ejemplos de normalización:

- Campañas Nocturnas -> Campanas Nocturnas
- Tracción, Dos Piernas Mejor que Una -> Traccion, Dos Piernas Mejor que Una
- ¿Ser o No Ser? Esa es la Cuestión -> Ser o No Ser Esa es la Cuestion

## Modificación de la interfaz (modo comparación)

La interfaz en el modo comparación mantendrá todos los elementos originales aunque como ya se ha mencionado se modificará el formato de las tablas y gráficas para albergar los datos de las obras que se comparan en lugar de sólo la del usuario.

La única ampliación significativa será una línea de información en donde se encuentran los datos del fichero actual. En ella se mostrará el nombre de la obra con la que se está comparando la del usuario y dos nuevos iconos al final de la misma: el icono pergamino y el icono equis.

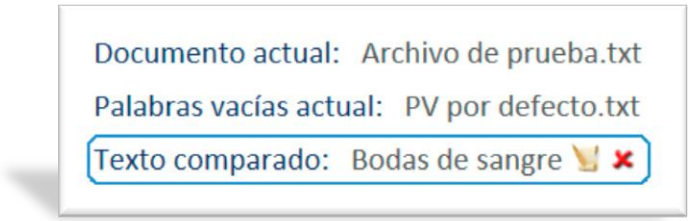



*Figura 9.18: Nueva línea de información y enlace de salida del modo de comparación*

**Icono pergamino:**

Situar el cursor sobre él desplegará un panel flotante con la información de la obra. La información mostrada será toda aquella que se haya incluido en el archivo XML de su colección.

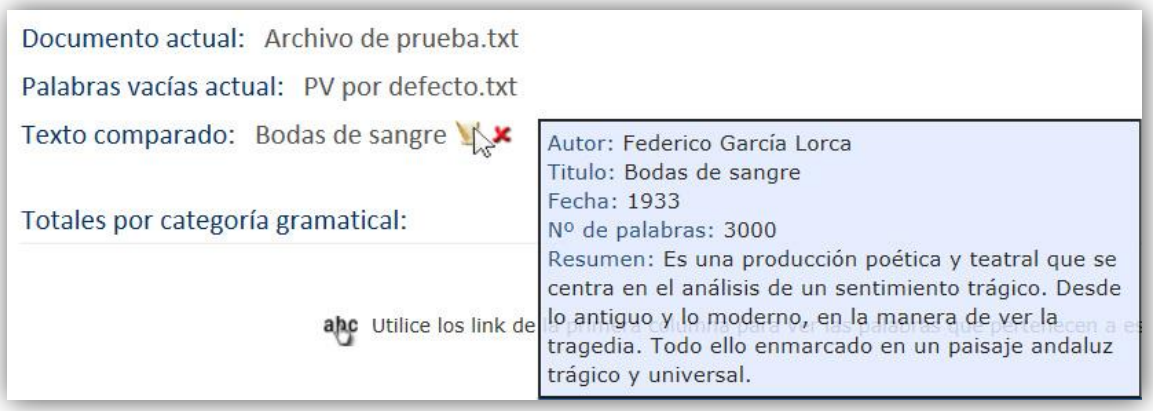



*Figura 9.19: Ventana desplegable con la información de la obra*

**Cancelar comparación:**

Hacer clic sobre el icono de la equis deberá devolver el programa al modo normal, aunque no deshará la comparación para facilitar poder pasar de un modo a otro rápidamente.

# 10 Implementación

En este capítulo se explican las características más relevantes de esta fase, sin embargo se ha optado por no incluir el código implementado aunque puede ser consultado en el DVD adjunto a esta memoria.

## *10.1 Clase silabear*

La primera clase en ser modificada ha sido la que tenía como objetivo separar las palabras en sílabas y remarcar la sílaba tónica. En primer lugar se realizó una llamada al servicio silabear del TIP que devolvía la información necesaria para realizar el trabajo de descomposición de la palabra.

La información que devuelve el servicio y el orden en el que lo hace es el siguiente:

1. *Primer dato:* Muestra el número de sílabas
2. *Número variable de datos:* que señalan las posiciones dentro de la palabra en las que empieza cada sílaba
3. *Último dato:* indica qué sílaba es la tónica

Con estos datos se debe generar una palabra cuyas sílabas estén separadas por guiones y la sílaba tónica se encuentre entre corchetes salvo que sea un monosílabo en cuyo caso se deja tal cual. El nuevo algoritmo comprende la creación de la ristra final en base a la información recibida y la petición y recepción de la información que ofrece el servicio.

| Palabra enviada | Información devuelta por el servicio | Ristra resultante |
|---|---|---|
| De | 111 | De |
| Esto | 2131 | [Es]-to |
| Esdrújula | 413682 | Es-[drú]-ju-la |
| Importante | 413693 | Im-por-[tan]-te |

*Tabla 10.1: Ejemplo de palabras enviadas, información devuelta y ristra resultante*

## 10.2 Nueva información n-gramas

El mayor problema fue el orden en el que se realizan las operaciones en el sistema original. Para poder extraer información de las palabras tienen que pasar también por un sistema de normalización, de forma que, por ejemplo la abreviatura U.L.P.G.C. pasará a ser ULPGC. Este proceso tiene lugar en la función encargada de fabricar las oraciones. Aunque lo ideal sería insertar una nueva función para los N-gramas posteriormente existe una funcionalidad adicional en el interior de la función de las oraciones que se encarga de seleccionar las palabras que ya hubieran sido analizadas con anterioridad y eliminarlas para que no fueran enviadas de nuevo a la fábrica de palabras (no era necesario fabricar dos veces la misma palabra). Al eliminarse palabras del interior de las oraciones no se pueden extraer los distintos N-gramas después de fabricarse las oraciones pero tampoco puede hacerse antes porque los N-gramas son subconjuntos de dichas oraciones y se requieren las palabras normalizadas.

Por esa razón fue necesario colocar el algoritmo en el interior de aquel que trata las oraciones y que se encuentra en la clase Analizador Gramatical en lugar de crearlo de forma independiente tal y como están hechos los de crear palabras y crear párrafos; así mismo eliminación de duplicidad a la hora de crear las palabras debía realizarse igualmente tras la creación de los n-gramas.

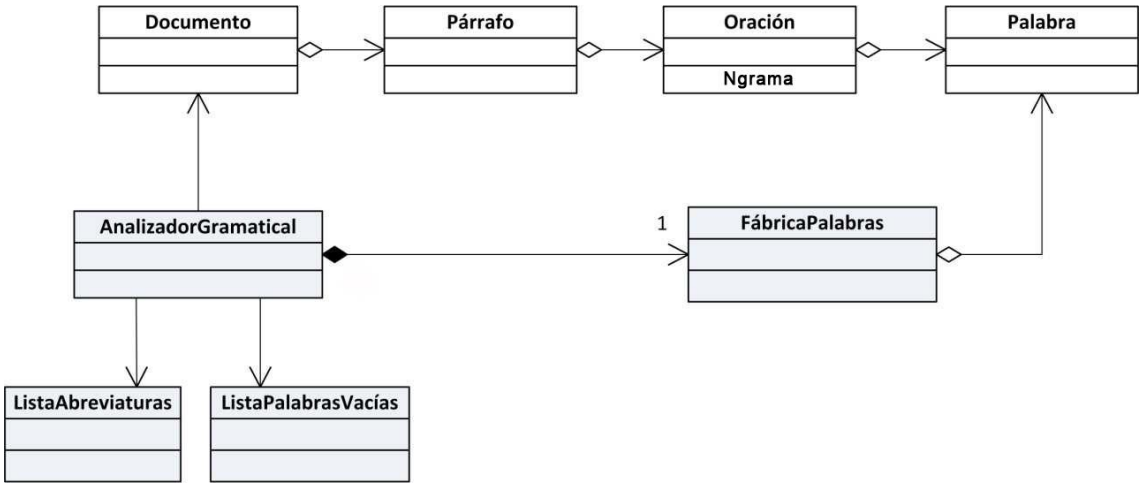



*Figura 10.1: Diagrama de clases del módulo Procesamiento*

Una vez integrado el nuevo algoritmo se crearon nuevos ficheros de serialización y se actualizó el modelo de enrutamiento para poder acceder a ellos.

## 10.3 Herramienta de comparación

La nueva herramienta se distribuyó mediante un sistema de carpetas y archivos XML y TXT puesto que no era necesario, en base a su volumen, la utilización de una base de datos, que por su lado requiere más trabajo y permite menos flexibilidad.

A continuación se muestra la jerarquía de carpetas utilizada:

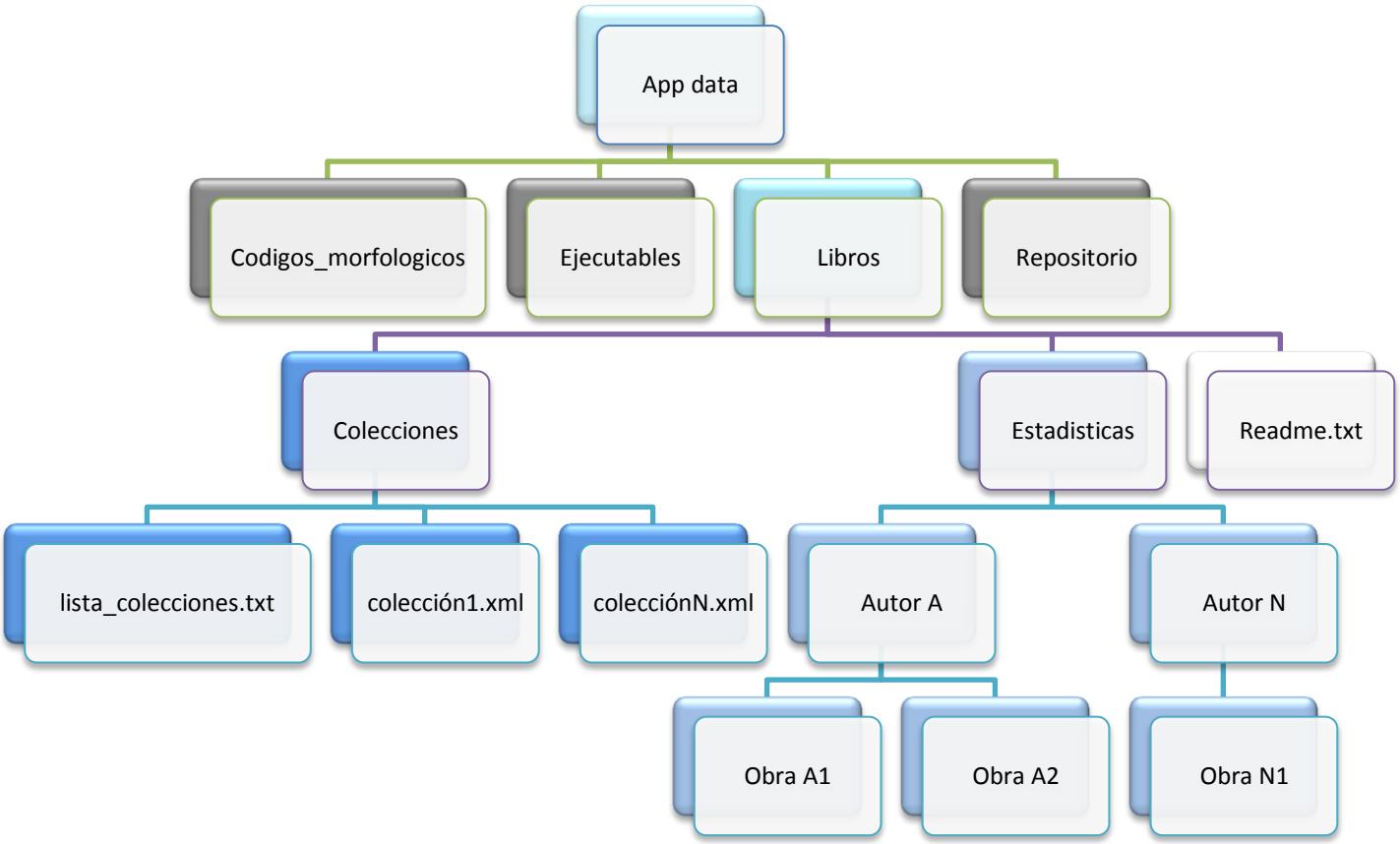



*Diagrama 10.1: Muestra la estructura de directorios y archivos para las obras almacenadas*

Gracias a este sistema es posible que un mismo libro muestre información distinta, según la colección en la que se encuentre, pues podría ser relevante para en unas pero no en otras y cada obra en una misma colección puede mostrar datos distintos si fuera necesario. Obras distintas con el mismo título no generan conflicto al estar situadas en las carpetas de sus respectivos autores y para conseguir la desambiguación cuando se muestran todas las obras a la misma vez se añade el nombre del autor al final de los títulos.

- **Carpeta App data:** Propia del sistema donde se almacenan datos necesarios para su correcto funcionamiento.

- **Carpeta Libros:** Carpeta que contiene las colecciones, estadísticas y el archivo de instrucciones

- **Archivo de texto Readme.txt:** Detallas las instrucciones necesarias para la correcta administración de la herramienta e incluye ejemplos de cada una de sus opciones para facilitar su manipulación.

- **Carpeta Colecciones:** incluye el fichero de texto que enlaza las colecciones que se muestran al usuario con el fichero XML que las define y los mencionados archivos XML asociados.

- **Carpeta Estadisticas:** Incluye las carpetas de los distintos autores cuyas obras han sido parametrizadas y almacenadas.

- **Archivo de texto lista_colecciones.txt:** Recoge las distintas colecciones que son mostradas al usuario y las enlaza con el archivo XML donde se encuentran las obras de le pertenecen.

- **Archivo XML colección1-N:** lista las distintas obra y la información que se muestra al usuario de las mismas; cada archivo XML pertenece a una colección distinta.

- **Carpeta Autor A-N:** Carpetas pertenecientes a los distintos autores que contienen las estadísticas de cada una de sus obras. No existe límite para el número de autores que pueden incluirse

- **Carpeta Obra A1-A2:** Carpetas con las estadísticas de las distintas obras de un autor, no hay límite de obras que pueden incluirse.

- **Carpeta Obra N1:** Carpeta que contiene las estadísticas de la obra de otro autor, los nombres de las obras de distintos autores pueden coincidir ya que se desambiguará mediante el nombre del autor.

**Nota:** Las instrucciones necesarias para el correcto funcionamiento de la herramienta así como ejemplos de su administración se encuentran tanto en el fichero Readme.txt mencionado anteriormente como en el apartado de Diseño en esta misma memoria.

## 10.4 Formato de las tablas

Para poder modificar las tablas se hizo uso de la herramienta CSS que ya incluía el proyecto inicial. Fue necesario crear números estilos e integrar su utilización en el código interno del programa para que se asignara dinámicamente según fuera necesario en lugar de hacerlo de forma fija en su totalidad.

Por ejemplo las columnas resaltadas en las tablas combinadas buscan en el encabezado la abreviación del autor con el que se está haciendo la comparación, pues es lo que determina si la columna pertenece a la obra de nuestro usuario o a una de las nuestras.

De esta forma no hay que definir de forma estática todas y cada una de las columnas que deben remarcarse, cosa que se complica por el alto número de columnas y que podría generar trabajo adicional por el hecho de que el programa pueda variar en el futuro. Independientemente de la posición en la que se sitúen las columnas de nuestras obras el sistema dinámico las localizará y resaltará para una fácil visualización.

## 10.5 Combinación de tablas

Para combinar las tablas fue necesario ampliar la estructura de los gridview a partir de los cuales se formaban y dado que no todas las tablas se modificaron de la misma manera hubo que modificarlos uno a uno.

| Con palabras vacías | Total | Total Lope de Vega |
|---|---|---|
| Caracteres | 95 | 112 |
| Palabras | 22 | 25 |
| Palabras diferentes | 17 | 21 |
| Oraciones | 2 | 3 |
| Párrafos | 2 | 2 |

*Figura 10.2: tabla con columna de totales añadido*

| Con palabras vacías | Media | Media Neruda | Desviación típica | Desviación típica Neruda |
|---|---|---|---|---|
| Caracteres por palabra | 4 | 4 | 2,53 | 3,03 |
| Caracteres por oración | 48 | 37 | 6,36 | 20,74 |
| Caracteres por párrafo | 48 | 56 | 6,36 | 0 |
| Palabras por oración | 11 | 8 | 1,41 | 3,51 |
| Palabras por párrafo | 11 | 12 | 1,41 | 0,71 |
| Oraciones por párrafo | 1 | 2 | 0 | 0,71 |

*Figura 10.3: tabla con columnas intercaladas añadidas*

| Nº Palabras | Nº Palabras Neruda | % Palabras | % Palabras Neruda | Nº Caracteres |
|---|---|---|---|---|
| 1 | 1 | 4,55 | 4 | 1 |
| 7 | 8 | 31,82 | 32 | 2 |
| 2 | 3 | 9,09 | 12 | 3 |
| 3 | 4 | 13,64 | 16 | 4 |
| 2 | 2 | 9,09 | 8 | 5 |
| 3 | 2 | 13,64 | 8 | 6 |
| 3 | 1 | 13,64 | 4 | 8 |
| 0 | 3 | 0 | 12 | 9 |
| 1 | 0 | 4,55 | 0 | 10 |
| 0 | 1 | 0 | 4 | 13 |

*Figura 10.4: tabla con datos combinados en base a una columna que también es un dato*

Aunque la tónica general era remarcar dinámicamente todas las tablas hubo una en particular que fue necesario marcar de forma estática pues el encabezado de su columna no requería de la abreviación del nombre del autor y no existía ningún canon que pudiera garantizar que no se marcaran otras que no debían.

| Forma canónica | Cat. gramatical | Obra origen | Etimologías | Acepciones | Antigua |
|---|---|---|---|---|---|
| a | prep. | Ambas | 2 | 23 | No |
| adelante | adv. lug. | Usuario | 1 | 3 | No |
| armado | v. pron. | Usuario | 1 | 27 | No |
| caso | v. tran. | Neruda | 3 | 1 | No |
| como | conj. sub. comp. | Neruda | 1 | 0 | No |
| correctamente | adv. modo | Neruda | 1 | 1 | No |

*Figura 10.5: tabla con datos combinados y metadatos*

## *10.6 Combinación de gráficas*

Algunas gráficas requerían que se modificara el origen de sus datos mientras que otras necesitaban datos adicionales. Uno de los mayores problemas fue el de los parámetros que el sistema original daba por sentado y que dejaban de cumplirse al realizarse la comparación.

Por ejemplo, el rango de las gráficas era siempre determinado por el archivo que contenía las palabras vacías, dado que aquel que no las contenía no era más que un subconjunto del primero y nada de lo que se cumpliera para el segundo dejaba de ser cierto en el primero. Sin embargo al realizar las comparaciones lo que se cumple para la del usuario (como por ejemplo el tamaño de la palabra más larga) no se cumplía necesariamente para la segunda obra que podía o no tener una palabra de mayor longitud.

Eso supuso un problema a la hora de determinar los valores mínimos y máximos que debían tener las gráficas, más aún, no fue hasta bien avanzado el proyecto que no se detectó el hecho de que algunos valores quedaban fuera del rango de las gráficas porque no se había estudiado en profundidad la forma en la que las gráficas definían sus tamaños. Finalmente mediante varias comparaciones las gráficas se establecen en base a los máximos y mínimos de ambas obras evitando que algunos valores quedaran ignorados.

Otro cambio a tener en cuenta era el valor de la leyenda que inicialmente eran valores fijos pero que ahora dependían tanto del nombre del archivo subido por el usuario como de la obra escogida para realizar la comparación. Esto produjo la necesidad de realizar de crear nuevas variables que hicieran las peticiones necesarias y almacenaran los valores para asignarlos cuando fuera preciso.

Las gráficas que tuvieron que ampliarse de una entrada de datos a dos quedaron no necesitaron modificaciones adicionales a las mencionadas

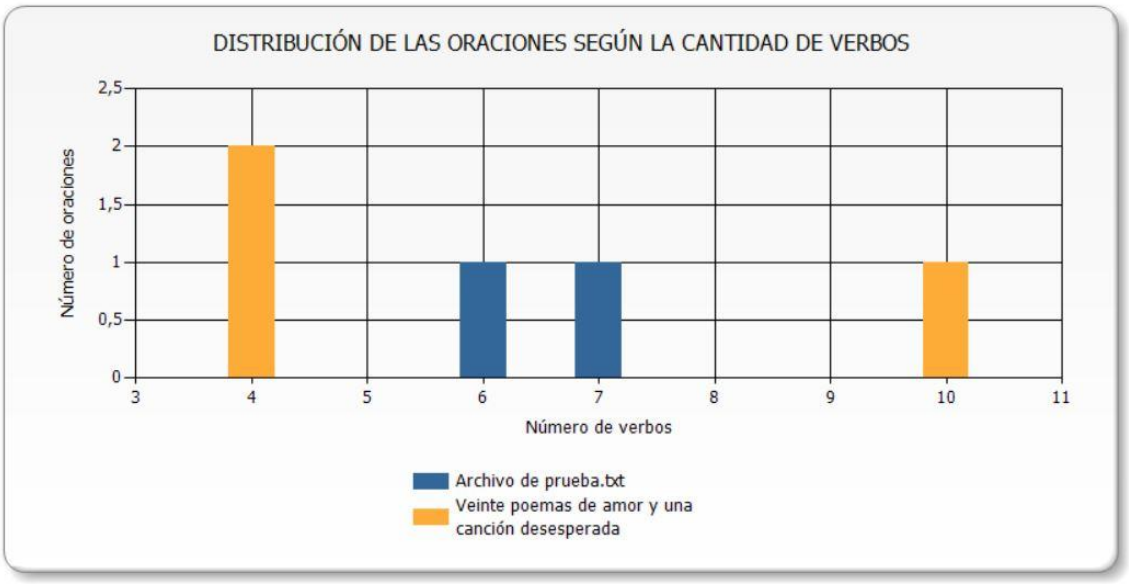



*Figura 10.6: gráfica que alberga una nueva entrada de datos*

Sin embargo aquellas que ya presentaban dos entradas necesitaron código adicional para evitar que quedaran saturadas. Se incluyó una nueva opción para ver las comparaciones en grupos de dos, los que eran relevantes entre sí. Mediante activación de una casilla la gráfica modifica la entrada de datos alterando dinámicamente su forma y permitiendo que se puedan comparar por separados los casos en los que se muestran las palabras vacías y en los que no. Dado que presenta más información se optó por presentar inicialmente las gráficas que incluyen las palabras vacías.

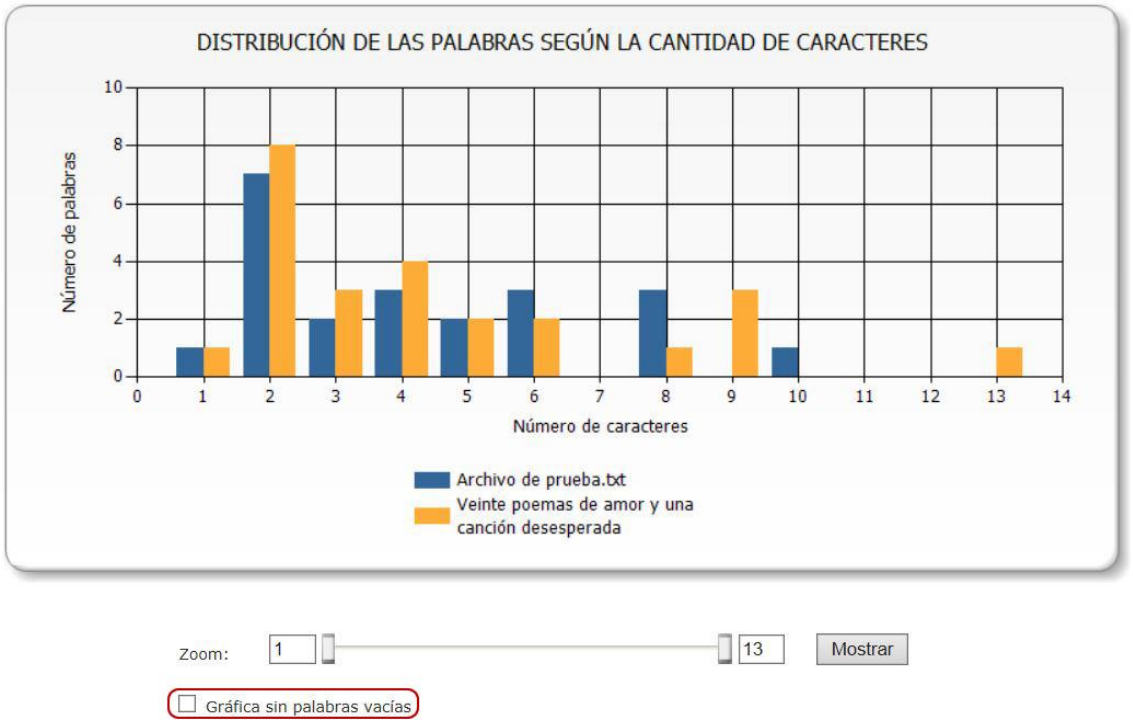



*Figura 10.7a: gráfica con palabras vacías*

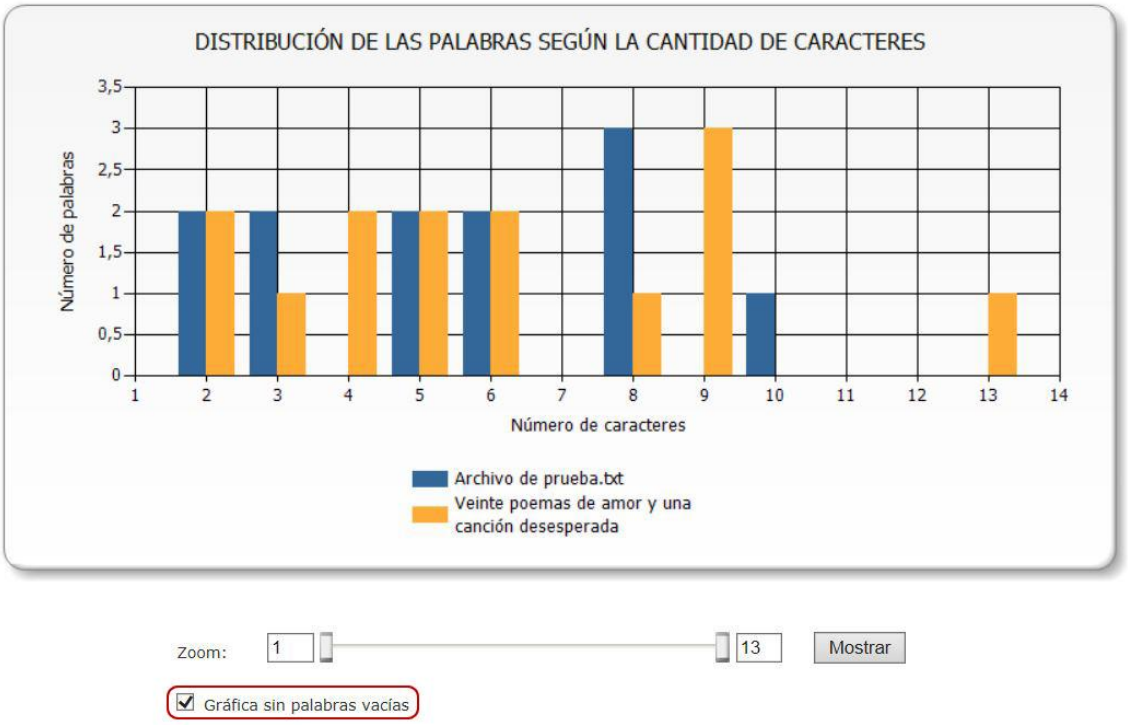



*Figura 10.7b: gráfica sin palabras vacías*

# 11 Pruebas

En esta fase del desarrollo se deben realizar multitud de pruebas para garantizar en la medida de lo posible el correcto funcionamiento del sistema. Aunque está situada al final, es una tarea que debe realizarse a lo largo de la fase de desarrollo pues detectar que se ha estado trabajando sobre clases, módulos o datos defectuosos podría tirar por la borda meses de trabajo. Es totalmente imposible garantizar que la aplicación está libre de errores en su totalidad pues hay tantas formas de usarla como textos existan. Sin embargo se ha realizado una amplia batería de pruebas para garantizar que los casos más comunes (y algunos de los que menos) no generar errores ni problemas de ejecución.

## *11.1 Pruebas de Recuperación*

Este tipo de pruebas intentan generar un error en el sistema y comprobar que es capaz de reponerse y continuar funcionando. En esta ampliación se detectaron numerosos errores de este tipo debido a un cambio en las listas usadas internamente por Paramtext que fue rápidamente subsanado por parte de los responsables de la herramienta.

La ampliación del sistema no presentaba casos especiales que pudieran dar lugar a inestabilidad debido a la acción de los usuarios. Debido a que no depende de la interacción de terceros se han podido subsanar los errores de este tipo con un alto grado de fiabilidad

## *11.2 Pruebas de seguridad*

Este sector comprueba que los mecanismos de seguridad incorporados impiden que el sistema sea vulnerado mediante la acción de actores externos. Debido a que la ampliación del sistema no permitía el envío adicional de datos con respecto a su antecesor no fue necesario realizarlas.

## *11.3 Pruebas de resistencia*

La carga añadida al proceso PMT que es el que consume recursos del servidor ha sido bastante controlada. Por un lado se le ha rebajado al añadir el servicio silabeador que realizar parte del trabajo mientras que por el otro se han añadido los ngramas que sí consumen algo de tiempo adicional. En el cómputo global Paramtext mantiene un tiempo de ejecución equivalente, quizás ligeramente superior que mantiene los mismos resultados obtenidos en las pruebas de la versión original: cuando aumenta el número de usuarios aumenta sensiblemente el tiempo de ejecución pero se siguen obteniendo los resultados sin problema.

## *11.4 Pruebas de rendimiento*

Este tipo de pruebas se basan en medir el rendimiento del software en tiempo de ejecución y requieren de instrumentación tanto software como hardware para los procesos de monitorización y medición. Al desarrollar este proyecto utilizando como entorno de desarrollo Microsoft Visual Studio, se ha decidido emplear la opción de generación de perfiles que ofrece la aplicación para este fin. Se trata de una herramienta de diagnóstico que recopila los datos de rendimiento de las aplicaciones, proporcionando información como el tiempo que se emplea en cada una de las funciones o la utilización de memoria por parte del sistema a lo largo del tiempo.

A través de la utilización de estas herramientas ha sido posible determinar que mientras que para el subsistema Paramtext, con el que los usuarios interactúan, no existen problemas destacables, el subsistema PMT encargado de realizar la parametrización morfológica de los textos posee ciertos inconvenientes a tener en cuenta. Se obtuvo que del total del tiempo necesario por el subsistema para completar la parametrización morfológica de un texto, el 85% del tiempo es destinado a obtener la información morfológica de las palabras, mientras que el otro 15% se destina al resto de actividades como la validación del archivo, su procesado y cálculo de resultados.

Si se establece que como término medio un fichero de texto de 100KB de tamaño es parametrizado en aproximadamente 42 segundos, significa que casi 36 de esos segundos son destinados únicamente a la obtención de la información morfológica de las palabras. Esto implica que el tiempo necesario por la aplicación para completar el proceso de parametrización y por tanto el tiempo que debe esperar el usuario para poder empezar a visualizar los resultados, está siempre determinado por el módulo de lematización, un hecho que es necesario asumir puesto que se utiliza un servicio web externo a este proyecto.

# 12 Resultados

Tras llevar a cabo todas las fases de desarrollo del software se ha conseguido ampliar satisfactoriamente el sistema. Se ha conseguido añadir nuevas funciones y ampliar la información que mostraba anteriormente, se le ha dotado de funcionalidad y se ha mantenido el tiempo y modo de funcionamiento para evitar un posible impacto negativos en quienes ya fueran usuarios de la herramienta. Aunque no se pretende incluir una por una todas las mejoras se realizará un resumen de las más significativas.

## *12.1 vocabulario*

El apartado dedicado al vocabulario ha pasado de tener una sola sección con dos tablas a tener tres secciones con siete tablas, lo que ha aumentado significativamente el volumen de información mostrado. Además de las dos tablas pertenecientes a las estadísticas se pueden encontrar, por un lado una nueva tabla de Léxico donde aparece información extraída en su mayor parte de la RAE y por el otro cuatro nuevas tablas donde se presentan los distintos ngramas de tamaños comprendidos entre dos y cinco, ambos incluidos. Adicionalmente se ha incluido el uso del servicio Silabeador en lugar de realizarse completamente en el interior de la aplicación.


Léxico

abc Sitúe el cursor del ratón sobre una categoría gramatical para verla sin abreviar

| Forma canónica | Cat. gramatical | Etimologías | Acepciones | Pos. de la acepción | Entr. de la acepción | Antigua | Aparece en la RAE |
|---|---|---|---|---|---|---|---|
| vamos | interj. | 1 | 0 | 0 | 0 | No | No |
| a | prep. | 2 | 23 | 1 | 23 | No | Sí |
| ver | sust. | 2 | 2 | 1 | 2 | No | Sí |
| si | conj. sub. cond. | 1 | 10 | 1 | 1 | No | Sí |
| este | pron. dem. | 1 | 0 | 0 | 0 | No | Sí |
| funcia | sust. | 1 | 0 | 0 | 0 | No | No |
| porque | conj. sub. causal | 1 | 2 | 1 | 1 | No | Sí |
| no | adv. neg. | 1 | 7 | 1 | 7 | No | Sí |
| el | art. det. | 1 | 1 | 1 | 1 | No | Sí |
| he | adv. | 1 | 1 | 1 | 1 | No | Sí |
| armar | v. pron. | 1 | 27 | 17 | 5 | No | Sí |
| e | sust. | 2 | 3 | 1 | 3 | No | Sí |
| importante | adj. | 1 | 2 | 1 | 2 | No | Sí |
| para | prep. | 1 | 13 | 1 | 13 | No | Sí |
| poder | v. intr. | 2 | 5 | 4 | 2 | No | Sí |
| seguir | v. tran. | 1 | 15 | 1 | 12 | No | Sí |
| adelante | adv. lug. | 1 | 3 | 1 | 2 | No | Sí |

Exportar

*Figura 12.1: Nueva tabla de Léxico en la sección de vocabulario*

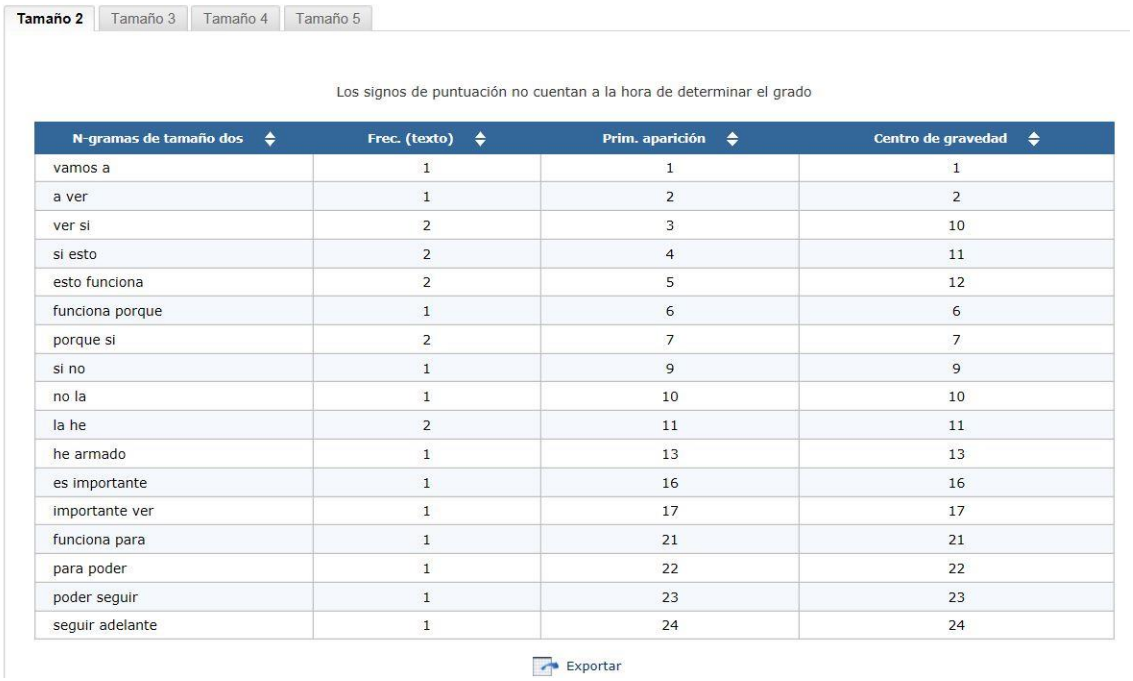



*Figura 12.2: Nuevas tablas de Ngramas separadas por las pestañas superiores en la sección de vocabulario*

## 12.2 Comparar

El nuevo apartado de comparación permite seleccionar mediante filtros (colecciones y autores) las obras con las que quiere realizarse una comparación. Incluye una descripción, instrucciones de uso, una sección de selección de obra y otra para mostrar su información.


*Figura 12.3: Nuevo apartado de la herramienta Comparar*

## 12.3 ParamTextComp

La nueva herramienta de comparación modifica completamente la interfaz de usuario para mostrar de forma clara y sencilla los datos combinados de las obras (una del usuario y otra de nuestro sistema). No se pretende incluir todas las tablas y gráficas afectadas cuyo número abarca treinta y siete (37) tablas y veintiuna (21) gráficas pero se pondrán algunos ejemplos de cada tipo.

| Con palabras vacías | Total |
|---|---|
| Caracteres | 95 |
| Palabras | 22 |
| Palabras diferentes | 17 |
| Oraciones | 2 |
| Párrafos | 2 |

| Con palabras vacías | Total | Total Lope de Vega |
|---|---|---|
| Caracteres | 95 | 112 |
| Palabras | 22 | 25 |
| Palabras diferentes | 17 | 21 |
| Oraciones | 2 | 3 |
| Párrafos | 2 | 2 |

*Figura 12.4: Tabla modo normal y modo comparación de totales perteneciente a Métrica - Informe*

| Con palabras vacías | Media | Desviación típica | Moda | Mediana |
|---|---|---|---|---|
| Caracteres por palabra | 4 | 2,53 | 2 | 4 |
| Caracteres por oración | 48 | 6,36 | - | 47 |
| Caracteres por párrafo | 48 | 6,36 | - | 47 |
| Palabras por oración | 11 | 1,41 | - | 11 |
| Palabras por párrafo | 11 | 1,41 | - | 11 |
| Oraciones por párrafo | 1 | 0 | - | 1 |

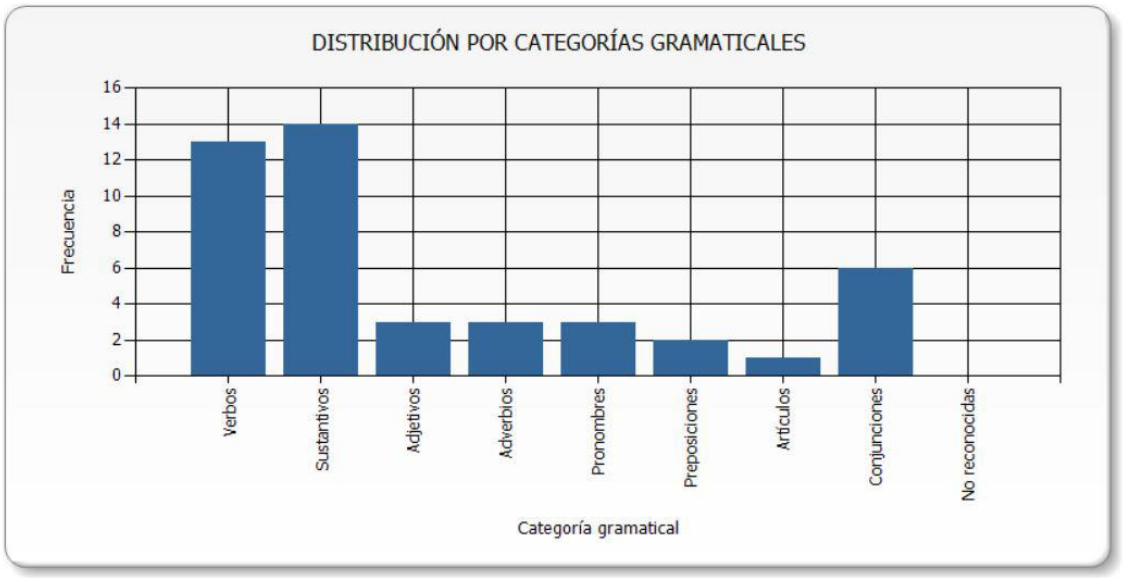
| Con palabras vacías | Media | Media Lope de Vega | Desviación típica | Desviación típica Lope de Vega | Moda | Moda Lope de Vega | Mediana | Mediana Lope de Vega |
|---|---|---|---|---|---|---|---|---|
| Caracteres por palabra | 4 | 4 | 2,53 | 3,03 | 2 | 2 | 4 | 4 |
| Caracteres por oración | 48 | 37 | 6,36 | 20,74 | - | - | 47 | 41 |
| Caracteres por párrafo | 48 | 56 | 6,36 | 0 | - | - | 47 | 56 |
| Palabras por oración | 11 | 8 | 1,41 | 3,51 | - | - | 11 | 8 |
| Palabras por párrafo | 11 | 12 | 1,41 | 0,71 | - | - | 11 | 12 |
| Oraciones por párrafo | 1 | 2 | 0 | 0,71 | - | - | 1 | 1 |

*Figura 12.5: Tabla modo normal y modo comparación de promedios perteneciente a Métrica - Informe*

| Forma canónica | Cat. gramatical | Etimologías | Acepciones | Pos. de la acepción | Entr. de la acepción | Antigua | Aparece en la RAE |
|---|---|---|---|---|---|---|---|
| vamos | interj. | 1 | 0 | 0 | 0 | No | No |
| a | prep. | 2 | 23 | 1 | 23 | No | Sí |
| ver | sust. | 2 | 2 | 1 | 2 | No | Sí |
| si | conj. sub. cond. | 1 | 10 | 1 | 1 | No | Sí |
| este | pron. dem. | 1 | 0 | 0 | 0 | No | Sí |

| Forma canónica | Cat. gramatical | Obra origen | Etimologías | Acepciones | Pos. de la acepción | Entr. de la acepción | Antigua | Aparece en la RAE |
|---|---|---|---|---|---|---|---|---|
| a | prep. | Ambas | 2 | 23 | 1 | 23 | No | Sí |
| adelante | adv. lug. | Usuario | 1 | 3 | 1 | 2 | No | Sí |
| armado | v. pron. | Usuario | 1 | 27 | 17 | 5 | No | Sí |
| caso | v. tran. | Lope de Vega | 3 | 1 | 1 | 1 | No | Sí |
| como | conj. sub. comp. | Lope de Vega | 1 | 0 | 0 | 0 | No | No |
| correctamente | adv. modo | Lope de Vega | 1 | 1 | 1 | 1 | No | Sí |

*Figura 12.6: Tabla modo normal y modo comparación de metadatos perteneciente a Vocabulario - Léxico*



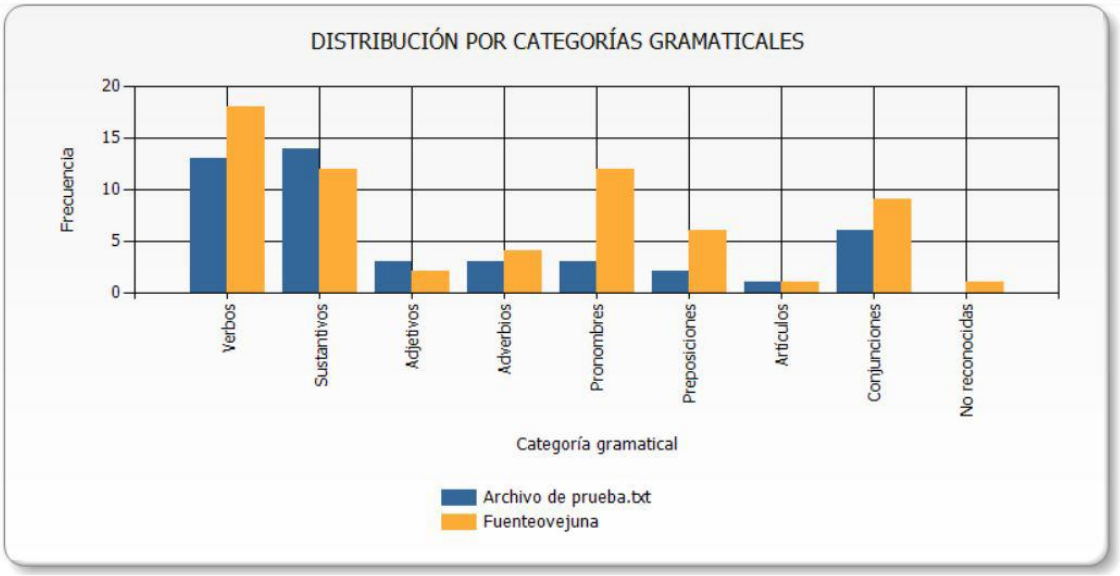
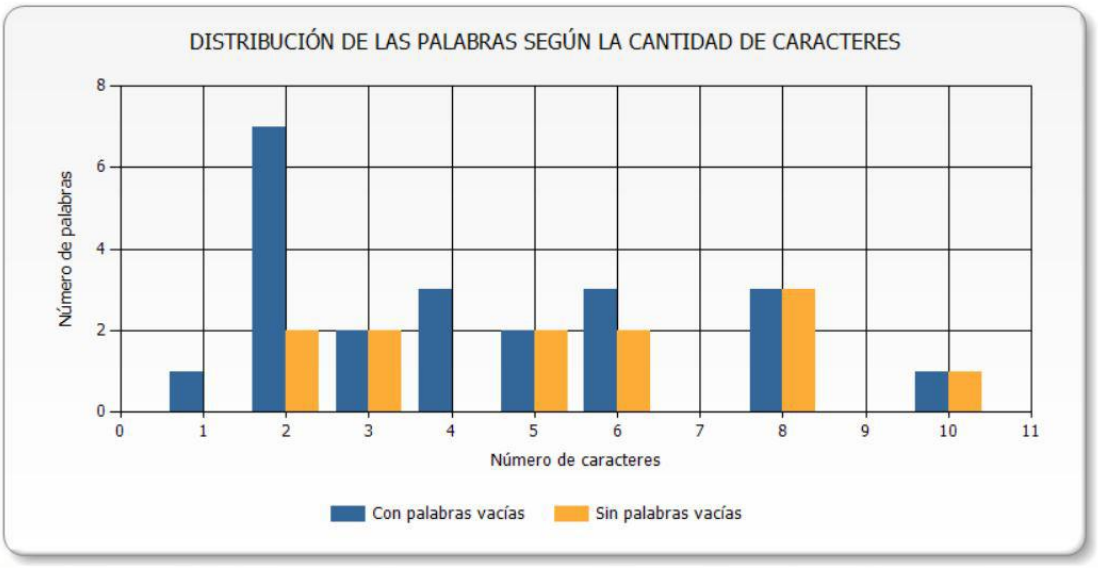
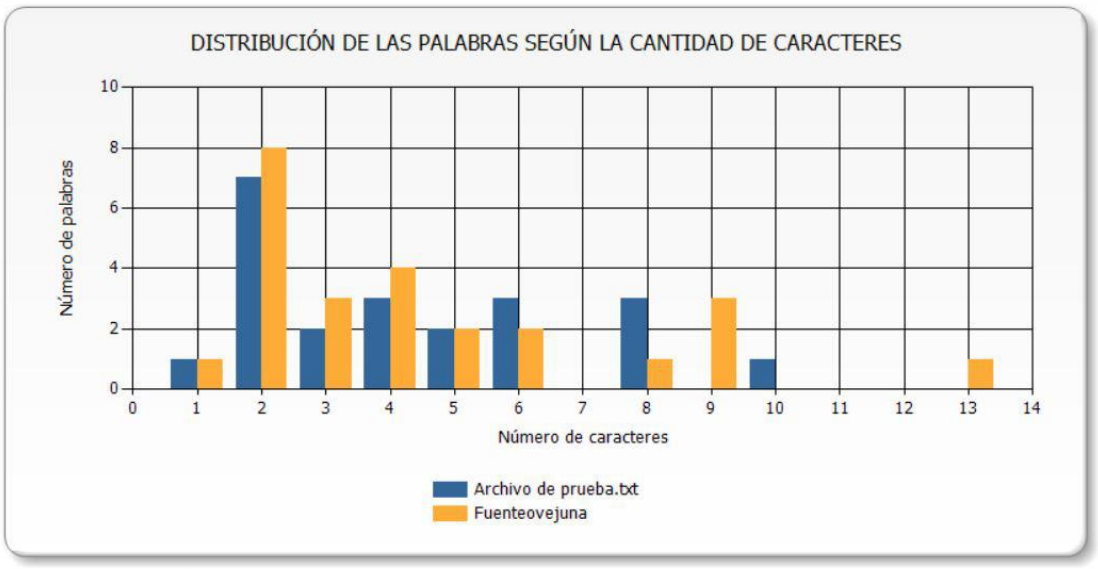
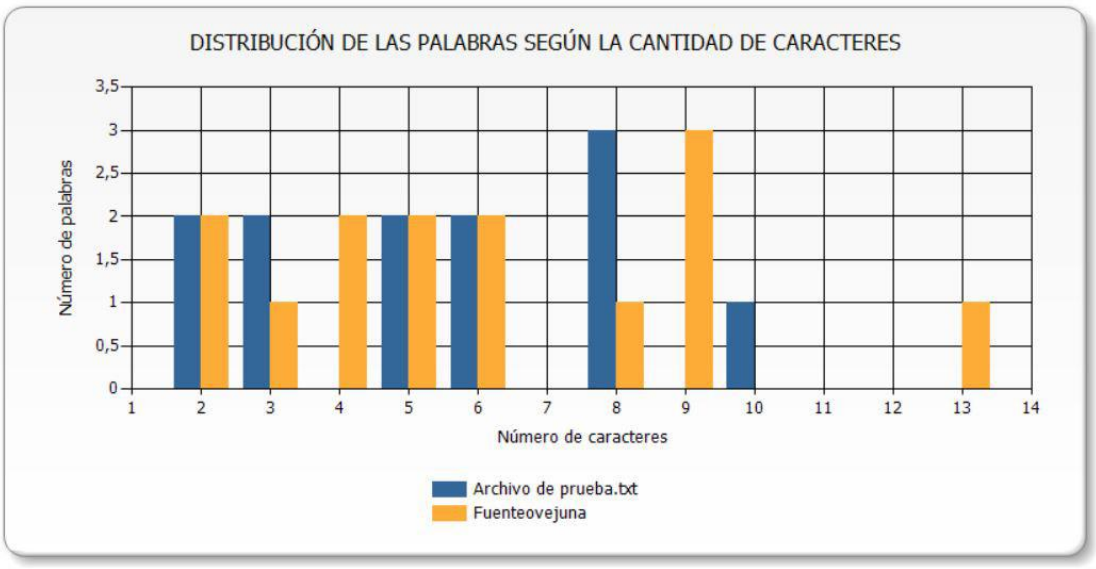




*Figura 12.7: Gráficas modo normal y modo comparación de una entrada a dos entradas de datos*




☐ Gráfica sin palabras vacías




☑ Gráfica sin palabras vacías

*Figura 12.8: Una gráfica con dos entradas de datos a dos graficas con cuatro entradas de datos (dos entradas en cada una)*

# 13 Conclusiones

Los resultados obtenidos han sido los esperados, un sistema adaptado para presentar una nueva funcionalidad que permite la comparación y con ella la capacidad de dar un significado, al menos relativo, a los valores que devuelve la herramienta. De igual forma se ha aumentado la cantidad de datos relativos al vocabulario, según la Real Academia de la Lengua, que es capaz de ofrecer al usuario y se han creado herramientas para poder mantener actualizados eso datos y se ha incluido información de los ngramas de tamaño dos a cinco. Por otro lado se ha cambiado el código incrustado que realizaba la separación de sílabas de una palabra por la llamada a un servicio externo y el posterior tratamiento de los datos. Por último se creó un sencillo sistema de ficheros que permite incrementar el número de obras disponibles para ser comparadas de forma sencilla y rápida, que además no requiere tocar el código del programa en ningún momento y gracias a ello tampoco es necesario detener la herramienta online para que surtan efecto.

El nuevo modo comparación presenta gráficas y tablas combinas en lugar de múltiples tablas simultáneas, lo que facilita la interpretación de los datos y ahorra espacio visual. Además el sistema puede pasar de ese punto al modo normal y viceversa con un simple clic de ratón y desde cualquier punto de la aplicación pues mantiene la ejecución de ambos modos de forma independiente.

En cuanto al rendimiento, el uso del servicio externo silabeador y la nueva información del vocabulario lo ha reducido ligeramente pero mejoras externas a este proyecto han conseguido que el tiempo global se mantenga. Gracias a que el modo comparación y normal se realizan paralelamente el paso de uno a otro se genera casi instantáneamente lo que van en consonancia con la fluidez de la que goza la aplicación.

El único que objetivo que quedó fuera del alcance de este proyecto fue la desambiguación, debido a problemas de compatibilidad y falta de tiempo, sin embargo la nueva funcionalidad de la aplicación le proporciona otro empujón más en la dirección correcta, aquella que la convierta en una herramienta de gran valor para todos los amantes y profesionales de la lengua escrita.

# 14 Trabajo futuro

Paramtext es una herramienta cada vez más funcional e innovadora en el área de la lengua escrita y el número de mejoras que pueden realizarse sobre él son muchas. Algunas de ellas ya se han mencionado con anterioridad a lo largo de la memoria; a continuación se enumeran las que actualmente pueden resultar más interesantes:

- Comparación entre dos archivos enviados por el usuario

Esta opción permitiría a un usuario comparar su obra no sólo con las que se han seleccionado en nuestras colecciones sino con otra que enviara también. La adaptación del sistema actual para incluir esta funcionalidad es, sin llegar a ser trivial, bastante sencilla y la única razón por la que no se incluyó en el proyecto actual fue la falta de tiempo.

- Incorporación de un desambiguador funcional

Este es quizás el caballo de batalla de la aplicación, su inclusión convertiría una herramienta útil de por sí en toda una referencia casi perfecta a la hora de comparar las estructuras de los textos. Debido a la gran dificultad y cantidad de trabajo que conllevaría crear uno propio se intentó hacer uso de uno externo pero problemas de compatibilidad lo dejaron fuera del proyecto aunque se hicieron grandes avances gracias al descubrimiento por parte del TIC de un servicio externo que podría usarse para ello en futuros proyectos.

- Automatización para blogs, foros o redes sociales

Actualmente la herramienta requiere que el usuario tome parte activa enviando y estudiando los resultados pero podría ser factible la ejecución automatizada en páginas web y su calificación posterior en base a parámetros de carácter subjetivo o no, basados en datos obtenidos en obras de renombre. Por ejemplo se podría realizar un estudio de los mejores artículos periodísticos para establecer algunas directrices que estructuralmente, y probablemente sin saberlo, siguieran todos ellos.

- Parametrización en otros idiomas

Si la aplicación logra un buen grado de aceptación por parte de los usuarios, se puede intentar conseguir un lematizador que permita obtener la información morfológica de palabras extranjeras, de esta manera sería viable utilizar el proyecto para parametrizar textos escritos en otros idiomas.

- Obtención de un índice de calidad

Fuera ya del ámbito informático, se propone tomar una muestra representativa de textos considerados de alta y baja calidad, para realizar un estudio científico de la parametrización morfológica de ellos. El objetivo es buscar similitudes y diferencias para intentar, si es posible, obtener un índice que permita determinar sin necesidad de leer un texto, si este es o no de calidad.

# 15 Conclusión personal

Es lógico pensar por quienes no tienen conocimientos de la carrera que aquellos que han acabado una ingeniería informática saben hacer cualquier cosa en lo referente a la informática; craso error, a medida que uno avanza en sus estudios se da cuenta de lo poco que sabe en particular y lo mucho que sabe en general. La carrera nos ha dado unas bases sobre las que trabajar, unos cimientos robustos y amplios en los que poder asentar todo lo que queramos, necesitemos o nos exijan para poder convertirnos en unos profesionales de aquellos ámbitos de la informática que vayamos a ejercer y disfrutar.

*No acabamos la carrera sabiendo "informática", lo hacemos preparados para poder aprenderla*.

# 16 Bibliografía

## 16.1 Libros


- [FPB03] J. Ferguson, B. Patterson, J. Beres. "La biblia de C#".
  Editorial: Anaya. 2003.

- [G01] J. A. González Seco. "El lenguaje de programación C#".
  Editorial: Ra-Ma. 2001.

- [S05] Ian Sommerville. "Ingeniería del software".
  Editorial: Pearson. 2005.

- [P02] R. Pressman. "Ingeniería del software. Un enfoque práctico".
  Editorial: Mc Graw-Hill. 2002.


## 16.2 Páginas web

El número de páginas webs consultadas para datos muy puntuales fue numerosa, las que se relatan a continuación son aquellas de las que se hizo un uso que podríamos definir como "intensivo".


- http://tip.dis.ulpgc.es

- http://msdn.microsoft.com/dn308572

- http://www.desarrolloweb.com/

- http://www.cristalab.com/tutoriales/4-html-css-y-javascript/

- http://forums.asp.net/

# 17 Anexo

## *17.1 Manual de usuario*

### 17.1.1 ¿QUÉ ES PARAMTEXT TIP?

ParamText TIP es una aplicación web destinada a la parametrización morfológica de textos, un conjunto de cálculos estadísticos basados en las características morfológicas del texto cuyos resultados se representan mediante la utilización de tablas y gráficas. Estos resultados permiten analizar la estructura del escrito desde el punto de vista morfológico y facilitan la comparación entre diferentes documentos.

### 17.1.2 ¿CÓMO PARAMETRIZAR UN DOCUMENTO?

La aplicación es muy sencilla de utilizar, sólo se necesita un documento cuyo tamaño y formato cumpla los requisitos que se establecen en la página de inicio de la aplicación (TXT, DOC, DOCX y PDF) y seguir los siguientes pasos:

***I. Enviar archivo***

En la sección habilitada para el envío de archivos, hacer clic en el botón "Examinar" para abrir la ventana de exploración de documentos. Buscar el archivo deseado y hacer doble clic sobre él. Automáticamente se iniciará el envío del archivo a nuestros servidores.

Una vez que el archivo haya sido recibido en nuestro sistema, se mostrará un cuadro con la información básica del fichero. En caso de querer cambiar el documento enviado, repita el proceso con el nuevo fichero. Si el archivo seleccionado no cumple los requisitos establecidos, se indicará mediante un aviso.

***II. Seleccionar lista de palabras vacías (opcional)***

El siguiente paso antes de comenzar el proceso de parametrización es seleccionar el tipo de lista de palabras vacías a utilizar. Dispone de dos opciones, utilizar la lista por defecto de nuestra aplicación, o bien enviar su propio listado de palabras vacías

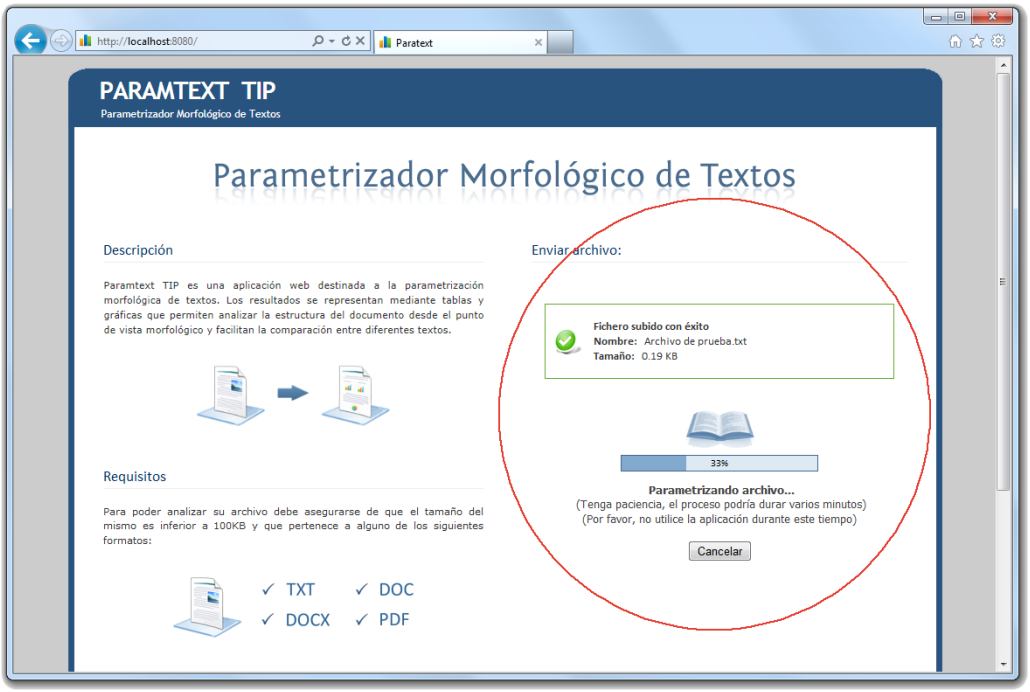
Si no comprende muy bien la utilidad de esta opción, no se preocupe, deje marcada la opción por defecto. Más adelante encontrará un apartado en el que se explica detalladamente en qué consisten las palabras vacías.

***III. Iniciar la parametrización***

Por último, sólo tiene que hacer clic sobre el botón "Parametrizar" y el sistema comenzará a procesar su documento. Puesto que este proceso puede llegar a durar minutos en función de las características del fichero enviado y del estado en el que se encuentre la red, se le ruega que tenga paciencia. No obstante, puede ver el estado en el que se encuentra el proceso mediante la barra de progreso que se proporciona, pudiendo cancelarlo en cualquier momento mediante la utilización del botón "Cancelar".

Para garantizar un flujo de ejecución correcto, evite utilizar la aplicación durante el tiempo en que se está realizando la parametrización de su archivo, en caso contrario, el sistema cancelará el proceso en curso.

Una vez finalizada la parametrización del documento, no será mostrado ningún mensaje, sino que será automáticamente redirigido a una de las páginas de resultados.




### 17.1.3 ¿CÓMO COMPARAR DOCUMENTOS? (v.2014)

La herramienta es muy sencilla de utilizar, sólo se necesita acceder a la zona de selección de obra mediante el enlace "Comparar" del menú superior de la página y seguir los siguientes pasos:

***I. Seleccionar colección***

De entre las distintas colecciones de obras ofertadas, seleccionar aquella que pueda contener el tipo de texto deseado. Se recomiendan aquellas que contengan un tipo de obra similar al que se ha enviado.

***II. Seleccionar el autor***

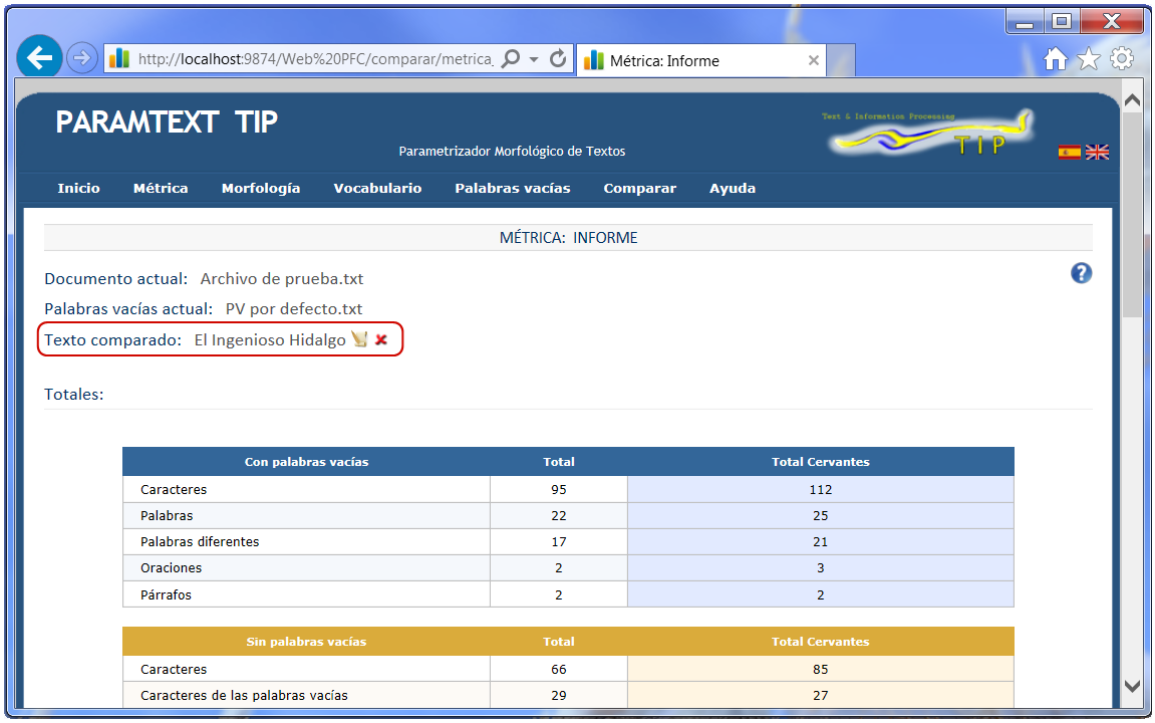
El siguiente paso antes de comenzar el proceso de comparación es seleccionar el Autor. Las obras están divididas según este criterio aunque existe la posibilidad de seleccionar "Todos" para que aparezca un listado completo.

***III. Seleccionar la obra e iniciar la comparación***

Por último, sólo tiene que seleccionar una obra y podrá ver una descripción de la misma y el botón de inicio de comparación en la parte inferior derecha de la página, al hacer clic sobre dicho botón la herramienta se pondrá en marcha, este proceso no debería tardar más que unos segundos.

Una vez finalizada la comparación del documento, no será mostrado ningún mensaje, sino que será automáticamente redirigido a una de las páginas de resultados. En ella podrá ver en todo momento el nombre de la obra con la que está realizando la comparación, y situados junto a ella, un icono (pergamino) que le permitirá ver los datos de la obra al colocar el cursor encima y otro icono (equis) que le permitirá deshacer la comparación haciendo clic sobre él.

Si se cancela una comparación, aparecerá una nueva línea de información indicando la última obra comparada y junto a ella un icono en forma de flecha que le permitirá rehacer la comparación en cualquier momento.


http://localhost:9874/Web%20PFC/comparar/metrica
Métrica: Informe
PARAMTEXT TIP
Parametrizador Morfológico de Textos
Inicio
Métrica
Morfología
Vocabulario
Palabras vacías
Comparar
Ayuda
MÉTRICA: INFORME
Documento actual: Archivo de prueba.txt
Palabras vacías actual: PV por defecto.txt
Texto comparado: El Ingenioso Hidalgo
Totales:
Con palabras vacías | Total | Total Cervantes
Caracteres | 95 | 112
Palabras | 22 | 25
Palabras diferentes | 17 | 21
Oraciones | 2 | 3
Párrafos | 2 | 2
Sin palabras vacías | Total | Total Cervantes
Caracteres | 66 | 85
Caracteres de las palabras vacías | 29 | 27


http://localhost:9874/Web%20PFC/metrica/informe_metrica.aspx
Métrica: Informe
PARAMTEXT TIP
Parametrizador Morfológico de Textos
Inicio
Métrica
Morfología
Vocabulario
Palabras vacías
Comparar
Ayuda
MÉTRICA: INFORME
Documento actual: Archivo de prueba.txt
Palabras vacías actual: PV por defecto.txt
Último texto comparado: Bodas de sangre
Totales:
Con palabras vacías | Total
Caracteres | 95
Palabras | 22
Palabras diferentes | 17
Oraciones | 2
Párrafos | 2
Sin palabras vacías | Total
Caracteres | 66
Caracteres de las palabras vacías | 29

### 17.2.4 PARTES DE LA APLICACIÓN

A continuación se detallan las principales partes del sistema:

***I. Menú***

Permite acceder a las opciones y resultados que proporciona la aplicación.

***II. Sección actual***

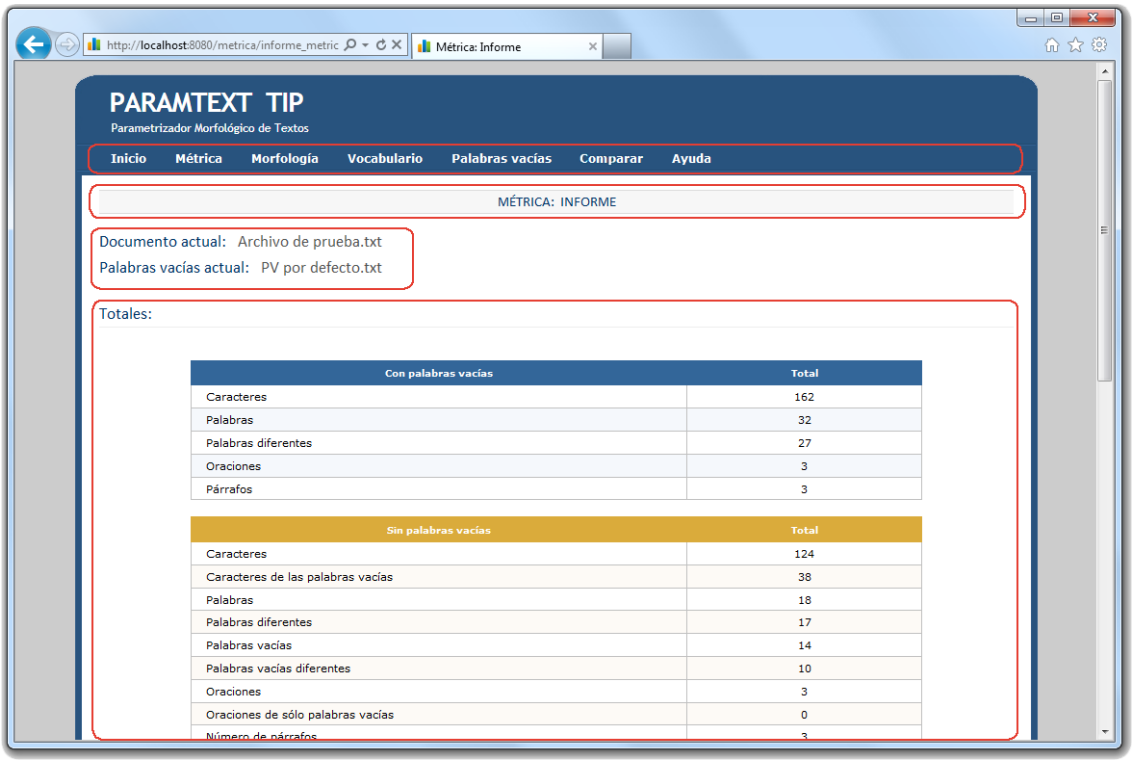
Muestra el nombre de la sección (resultado) que se está visualizando.

***III. Documentos actuales***

Indica el nombre del fichero y lista de palabras vacías actuales. Permite saber en todo momento a que documento se corresponden los resultados mostrados.

***IV. Bloque principal***

Es la sección más importante, ya que en ella se presentarán los resultados solicitados.



### 17.1.5 DISTRIBUCIÓN DEL MENÚ

El menú ha sido dividido en varias secciones para clasificar los resultados por categorías y facilitar así su localización. A continuación, se comenta el tipo de resultados que podrá consultar en cada una de sus secciones:

***I. Inicio***

Permite acceder a la página de inicio de la aplicación.

***II. Métrica***

Ofrece los resultados relacionados con la métrica del documento analizado. Podrá consultar por ejemplo: el número de palabras, el número de palabras diferentes, el número de oraciones, el número de párrafos, la cantidad de oraciones por párrafo, o bien el promedio, moda y mediana de palabras por oración o párrafo entre otros muchos datos.

***III. Morfología***

Proporciona una amplia información sobre las características morfológicas del texto. Podrá consultar por ejemplo: el número de palabras que pertenecen a cada categoría gramatical (verbos, sustantivos, adjetivos,…), el número de palabras según su flexión, o el promedio, moda y mediana de palabras de una determinada categoría gramatical que hay por oración o párrafo.

***IV. Vocabulario (v. 2014)***

Permite consultar la lista de palabras que aparecen en el texto analizado. Además, para cada una de ellas podrá visualizar su frecuencia de aparición, la posición del texto en la que aparece por primera vez, su información morfológica, categoría gramatical, numero de acepciones en la RAE, etc. También se ha dedicado un apartado para los distintos Ngramas existentes (de entre 2 y 5 palabras).

***V. Palabras vacías***

Esta sección permite cambiar la lista de palabras vacías a utilizar. También proporciona la posibilidad de descargar la lista de palabras vacías por defecto o la utilizada por el usuario en la última parametrización realizada.

***VI. Comparar (v.2014)***

Permite realizar una comparación entre el texto enviado y uno de los que se mantienen almacenados en nuestra base de datos. Dicha herramienta reestructurará por completo el Parametrizador para mostrar en todo momento y de forma clara los datos de ambas obras, facilitando así cualquier tipo de comparación.

***VII. Ayuda***

Permite acceder al documento de ayuda online de nuestra aplicación.

### 17.1.6 PALABRAS VACÍAS

Las palabras vacías (en inglés, stopwords) son aquellas que no tienen un significado propio y que por tanto no aportan ningún contenido semántico al texto. Es por ello que en muchas actividades relacionadas con el procesamiento de textos, este tipo de palabras tengan un tratamiento particular. Un ejemplo, son los motores de búsquedas, que descartan este tipo de palabras para ofrecer mejores resultados a los usuarios.

En nuestro caso, hemos decidido ofrecer los resultados de la sección Métrica desde dos perspectivas diferentes:

***I. Con palabras vacías***

Los resultados se ofrecen teniendo en cuenta todas las palabras del texto, por tanto las palabras vacías son contabilizadas.

***II. Sin palabras vacías***

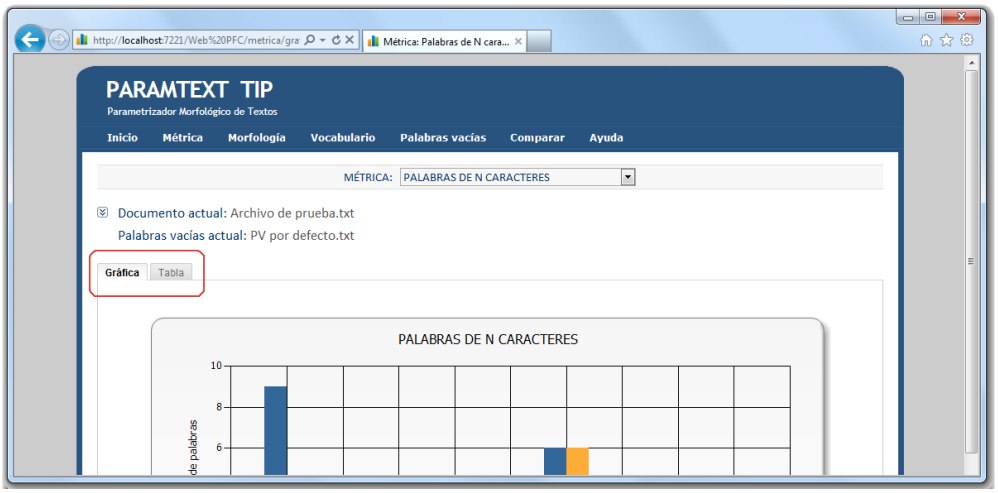
Los resultados se ofrecen sin tener en cuenta las palabras vacías, estas no serán consideradas y por tanto, tampoco contabilizadas.

ParamText TIP utiliza una lista de palabras vacías por defecto. En ella se han incluido las palabras vacías más frecuentes del español. Sin embargo, puesto que no existe un estándar que defina con exactitud cuáles son las palabras vacías del español, se ofrece también la posibilidad de que sea el usuario quien establezca su propia lista de palabras vacías.

Esta opción permitirá a los usuarios obtener resultados más precisos, ya que podrán descartar todas aquellas palabras que deseen. En la sección del menú "Palabras vacías", encontrará las pautas a seguir para poder confeccionar y utilizar su propio listado.

### 17.1.7 FORMATO DE LOS RESULTADOS

Los resultados proporcionados por ParamText TIP, pueden ser visualizados en dos formatos diferentes. Uno de ellos es en forma de gráfica, que es el formato mostrado por defecto. La otra posibilidad, consiste en visualizar la información en modo tabular. Puede alternar entre ambas vistas mediante la utilización de las pestañas habilitadas para ello.




***Gráficas***

Permiten visualizar los datos que conforman el resultado seleccionado en formato de gráfica de barras. Para facilitar la interpretación de la información mostrada, cada gráfica va acompañada de una leyenda y los títulos de sus correspondientes ejes.

En la parte inferior de la gráfica, encontrará un conjunto de opciones que permitirán cambiar ciertos parámetros de visualización. Las opciones disponibles son:

**I. Zoom**

Permite seleccionar un rango de valores específicos del eje X para visualizarlo con más detalle.

**II. Mostrar serie**

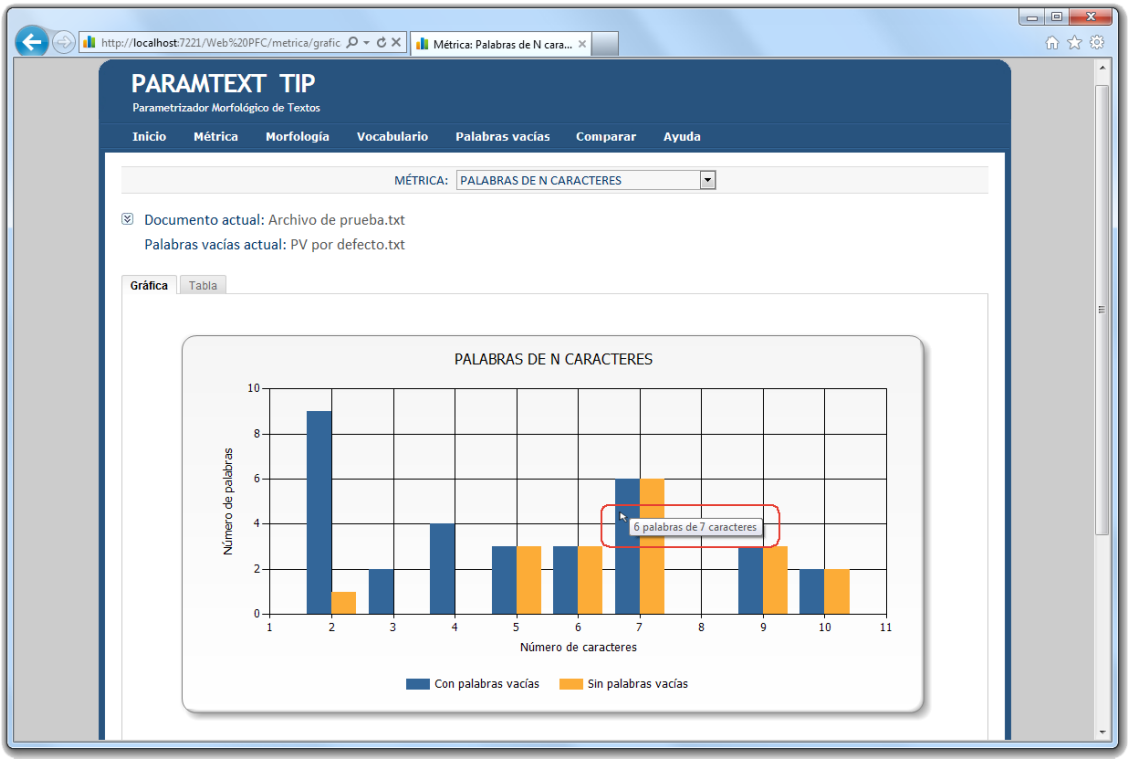
Por defecto, para facilitar la comparación de los resultados, las gráficas muestran simultáneamente los datos "Con palabras vacías" y "Sin palabras vacías". Esta opción permite restringir la visualización de la gráfica a una única serie.

**III. 3D**

Permite cambiar el modo de visualización de la gráfica a tres dimensiones.

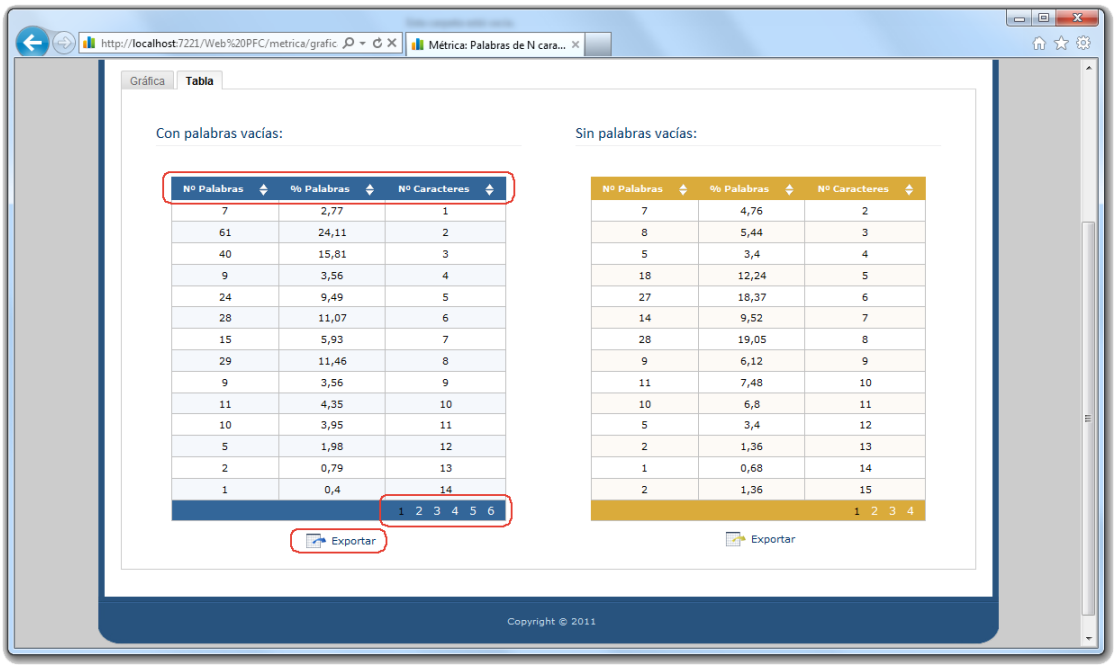
**IV. Tooltip**

Permite conocer los valores de los ejes X e Y que conforman cada una de las barras de la gráfica. Para ello, sólo hay que pasar el cursor del ratón sobre alguna de las barras y se mostrará la información correspondiente.



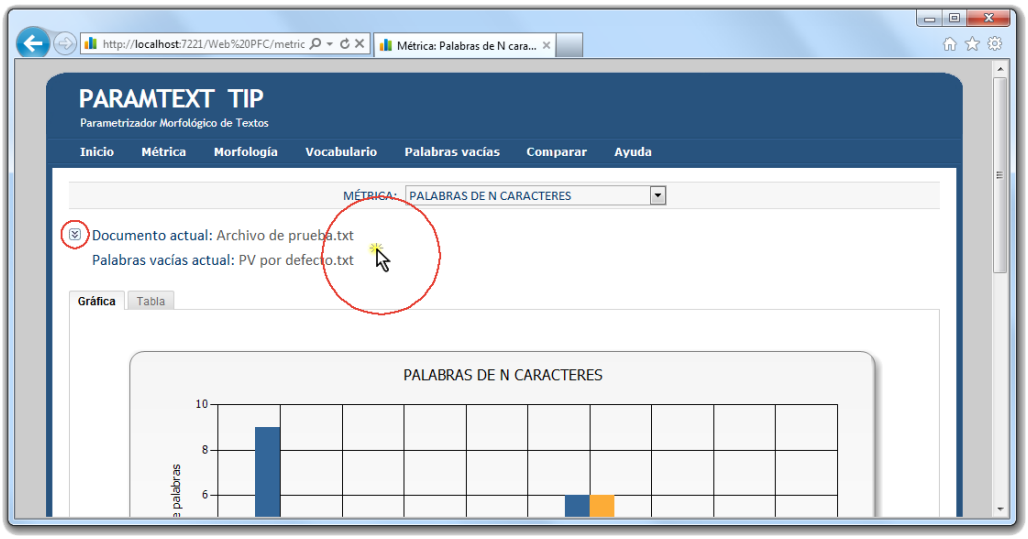

## Tablas

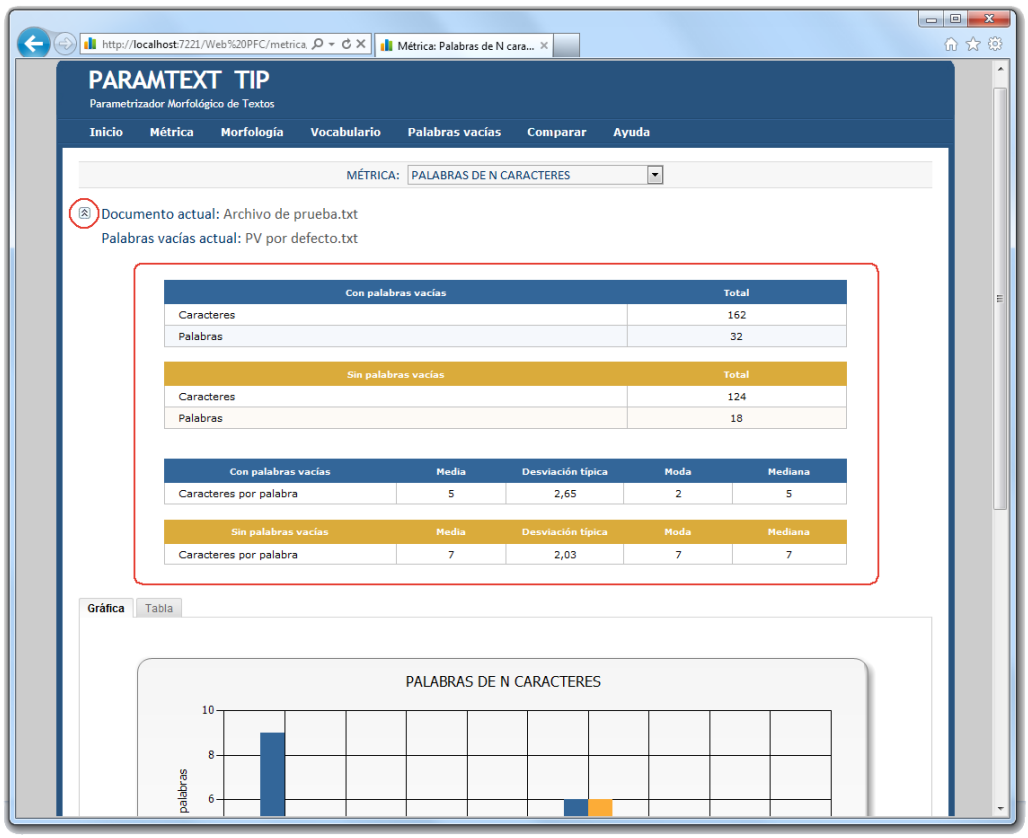
Ofrecen la misma información que las gráficas pero en formato tabular. Puesto que la mayor parte de los resultados implican el manejo de un gran volumen de datos, se ha decidido habilitar la paginación y ordenación por columnas de las tablas, facilitando así, la búsqueda y legibilidad de los datos. Además, mediante la utilización del botón "Exportar" que encontrará bajo cada una de ellas, podrá exportar el contenido de la tabla a un archivo de Excel.



Otra de las opciones disponibles cuando se consulta un resultado, ya sea en formato gráfico o tabular, es la visualización de un pequeño informe, que se pliega y despliega al hacer clic con el ratón sobre los datos del documento actual.

Este pequeño informe, contiene únicamente algunos de los resultados de los informes principales, que pueden ser útiles al usuario en función del resultado que está visualizando. Por ejemplo, en las imágenes inferiores, en las que el usuario está consultando como se distribuyen las palabras del texto según su tamaño en caracteres, el informe desplegable muestra como datos relevantes, el total de palabras y el total caracteres del texto. De esta manera, si el usuario deseara consultar esta información, no tendría que irla a buscar a la página del informe principal, sino que podría consultarla desde la misma página en la que se encuentra.



### 17.1.8 RESULTADOS PROPORCIONADOS

En este apartado se explican las características y utilidad de los principales resultados proporcionados por ParamText TIP:

***Informe métrica***

En esta sección se proporcionan resultados como:

- Número total de: caracteres, palabras, palabras diferentes, oraciones o párrafos.
- Promedio, desviación típica, moda y mediana de: caracteres por palabra, caracteres por oración, caracteres por párrafo, palabras por oración, palabras por párrafo u oraciones por párrafo.

Todos estos datos permiten hacerse una idea de cómo se estructura el texto en función de sus partes, como por ejemplo, si predominan palabras u oraciones cortas o largas. Estos resultados, al ser de carácter unitario, sólo se representan en formato tabular. Esta sección está disponible en: "Métrica → Informe".

***Distribución de las palabras según la cantidad de caracteres***

Muestra cómo se distribuyen las palabras del texto según la cantidad de caracteres. Es decir, indica el número de palabras cuya longitud en caracteres es N, donde N = {1, 2, 3,…}. Permite saber si las palabras del texto se caracterizan por ser pequeñas o grandes, e incluso detectar la cantidad de palabras que superan un determinado tamaño. Este resultado está disponible en: "Métrica → Gráficas → Palabras de N caracteres".

***Distribución de las oraciones según la cantidad de caracteres***

Muestra cómo se distribuyen las oraciones del texto según la cantidad de caracteres. Es decir, indica el número de oraciones cuya longitud en caracteres es N, donde N = {1, 2, 3,…}. Este resultado está disponible en: "Métrica → Gráficas → Oraciones de N caracteres".

***Distribución de los párrafos según la cantidad de caracteres***

Muestra cómo se distribuyen los párrafos del texto según la cantidad de caracteres. Es decir, indica el número de párrafos cuya longitud en caracteres es N, donde N = {1, 2, 3,…}. Este resultado está disponible en: "Métrica → Gráficas → Párrafos de N caracteres".

***Distribución de las oraciones según la cantidad de palabras***

Muestra cómo se distribuyen las oraciones del texto según la cantidad de palabras. Es decir, indica el número de oraciones que están formadas por N palabras, donde N = {1, 2, 3,…}. Este resultado está disponible en: "Métrica → Gráficas → Oraciones de N palabras".

***Distribución de los párrafos según la cantidad de palabras***

Muestra cómo se distribuyen los párrafos del texto según la cantidad de palabras. Es decir, indica el número de párrafos que están formados por N palabras, donde N = {1, 2, 3,…}. Este resultado está disponible en: "Métrica → Gráficas → Párrafos de N palabras".

***Distribución de los párrafos según la cantidad de oraciones***

Muestra cómo se distribuyen los párrafos del texto según la cantidad de oraciones. Es decir, indica el número de párrafos que están formados por N oraciones, donde N = {1, 2, 3,…}. Al consultar este resultado, puede surgir la duda de por qué se proporcionan los datos tanto desde el punto de vista "Con palabras vacías" como "Sin palabras vacías", cuando se supone que el número de oraciones del texto siempre son las mismas. La razón es muy sencilla, aunque es bastante raro, puede suceder que una oración esté formada únicamente por palabras vacías, por tanto, desde el punto de vista "Sin palabras vacías", el texto tendrá una oración menos. Este resultado está disponible en: "Métrica → Gráficas → Párrafos de N oraciones".

***Distribución por frecuencia***

Indica cómo se distribuyen las palabras del texto en función del número de veces que aparecen en él. Permite saber cuantas palabras del texto aparecen con poca o mucha frecuencia. Este resultado está disponible en: "Métrica → Gráficas → Distribución por frecuencia".

***Distribución por centro de gravedad***

El centro de gravedad de una palabra es la media de posiciones en las que aparece dicha palabra en el texto y por tanto proporciona una idea de la zona del texto en la que más o menos aparece dicha palabra con más frecuencia. Este resultado muestra como se distribuyen las palabras según su centro de gravedad y permite saber, por tanto, si las palabras del texto se concentran especialmente en alguna zona particular del texto. Este resultado está disponible en: "Métrica → Gráficas → Distribución por centro de gravedad".

**Distribución por primera aparición**

Este resultado muestra como se distribuyen las palabras según la posición de su primera aparición en el texto, por tanto permite detectar si a medida que avanza el texto aparecen nuevas palabras o se repiten las ya utilizadas. Este resultado está disponible en: "Métrica → Gráficas → Distribución por primera aparición".

**Distribución en el corpus**

Este resultado permite saber con qué frecuencia se utilizan en el español las palabras del texto. Después de haber realizado un estudio sobre la frecuencia de aparición de cada una de las palabras de un corpus, formado por 320575144 palabras, de las cuales, 309734 eran diferentes, se ha decidido dividir las palabras del texto en seis grupos:

- Frecuencia muy alta
  Son las palabras del texto cuya frecuencia en el corpus es superior o igual a 30000. Este tramo lo componen 939 palabras del corpus.

- Frecuencia alta
  Son las palabras del texto cuya frecuencia en el corpus es inferior a 30000, pero superior o igual a 2000. Este tramo lo componen 10087 palabras del corpus.

- Frecuencia media
  Son las palabras del texto cuya frecuencia en el corpus es inferior a 2000, pero superior o igual a 300. Este tramo lo componen 26295 palabras del corpus.

- Frecuencia baja
  Son las palabras del texto cuya frecuencia en el corpus es inferior a 300, pero superior o igual a 20. Este tramo lo componen 80264 palabras del corpus.

- Frecuencia muy baja
  Son las palabras del texto cuya frecuencia en el corpus es inferior a 20, pero superior a cero. Este tramo lo componen 192149 palabras del corpus.

- Frecuencia cero
  Son las palabras del texto que no aparecen ninguna vez en el corpus.

Por tanto, este resultado permite saber en base al corpus utilizado, qué cantidad de palabras del texto analizado, se consideran de uso frecuente o poco frecuente en el español. Este resultado está disponible en: "Métrica → Gráficas → Distribución en el corpus".

**Vocabulario**

Esta sección permite ver el listado de palabras que aparecen en el texto. La información se ha dividido entre las siguientes secciones:

***I. Estadísticas***

- Sílabas de la palabra
  Se muestran las sílabas separadas por guiones y entre corchetes se marca la sílaba tónica.

- Frecuencia en el texto
  Es el número de veces que aparece la palabra en el texto.

- Primera aparición
  Es la posición del texto en la que aparece la palabra por primera vez.

- Centro de gravedad
  Como se explicó anteriormente, el centro de gravedad de una palabra es la media de posiciones en las que aparece una palabra en el texto y por tanto nos da una idea de la zona del texto en la que más o menos aparece dicha palabra con más frecuencia.

- Frecuencia en el corpus
  Es el número de veces que aparece la palabra en el corpus analizado, cuyas características fueron explicadas en el apartado anterior. Este dato permite saber, si la palabra es utilizada o no con frecuencia en el español.

- Palabra invertida
  La palabra escrita al revés. Esta columna permite ordenar por el final de las palabras.

Otra característica importante de esta sección, es que proporciona la posibilidad de visualizar la información morfológica de cada una de las palabras. Para ello sólo hay que situar el cursor del ratón sobre la palabra cuya información morfológica se desea visualizar. Recuerde que debido a las características del lenguaje español, una palabra puede disponer de varias interpretaciones morfológicas. Este resultado está disponible en: "Vocabulario".

***II. Léxico (v.2014)***

- Forma canónica
  Indica la forma canónica de las palabras que aparecen en el texto.

- Categoría gramatical
  Muestra de forma abreviada la categoría gramatical a la que pertenece la palabra. Situar el cursor sobre su descripción muestra la categoría gramatical sin abreviar
- Etimologías
  Indica el número de etimologías que tiene la palabra.

- Acepciones
  Indica el número total de acepciones que presenta la palabra en el diccionario de la RAE.

- Posición de la acepción
  Indica la posición en la que se encuentra su categoría gramatical dentro de su definición en el diccionario de la RAE.

- Entradas de la acepción
  Indica el número total de entradas de su categoría gramatical en el diccionario de la RAE.

- Antigua
  Determina si la palabra se considera en desuso o muy antigua según el diccionario de la RAE.

- Aparece en la RAE
  Indica si dicha palabra se encuentra en el diccionario de la RAE.

***III. Ngramas (v.2014)***

En esta sección se muestran cuatro pestañas distintas, que se corresponden con los distintos tamaños de Ngramas disponibles.

- Frecuencia en el texto
  Es el número de veces que aparece el Ngrama en el texto.

- Primera aparición
  Es la posición del texto en la que aparece en Ngrama por primera vez.

- Centro de gravedad
  El centro de gravedad de un Ngrama es la media de posiciones en las que aparece en el texto y por tanto nos da una idea de la zona del texto en la que más o menos aparece dicho Ngrama con más frecuencia.

**Informe morfología**

En esta sección se proporcionan resultados como:

- Número total de palabras por categoría gramatical (verbos, sustantivos, adjetivo, adverbios, pronombres, preposiciones, artículos,…).
- Promedio, desviación típica, moda y mediana por oración y párrafo de: verbos por oración, verbos por párrafo, sustantivos por oración, sustantivos por párrafo, adjetivos por oración, adjetivos por párrafo,…
- Número total de palabras por flexión verbal (infinitivos, gerundios,…).
- Número total de palabras por flexión no verbal (singular, plural,…).

Todos estos datos permiten obtener una idea de cómo está estructurado el texto en función de sus características morfológicas, pudiendo determinar qué categorías gramaticales o flexiones predominan en él. Estos resultados al ser de carácter unitario sólo se representan en formato tabular.

Otro tipo de resultado que es posible visualizar desde esta sección, es obtener una lista de las palabras del texto que pertenecen a una determinada categoría gramatical o flexión. Para ello, sólo debe dirigirse a la tabla correspondiente y hacer clic con el ratón sobre la categoría gramatical o flexión deseada (Figuras 17.25 y 17.26). Al igual que en la sección "Vocabulario", situando el cursor del ratón sobre cualquiera de las palabras de la lista resultante, puede acceder a su información morfológica. Esta sección está disponible en: "Morfología → Informe".

***I. Palabras no reconocidas***

Permite examinar la lista de palabras que no han sido reconocidas morfológicamente por la aplicación. Esta aplicación actualmente sólo funciona para textos en español, por lo que si el texto contiene alguna palabra en otro idioma, también aparecerá en esta lista. Este resultado está disponible en: "Morfología → Palabras no reconocidas".

***II. Categorías gramaticales***

Indica cómo se distribuyen las palabras del texto según su categoría gramatical (verbos, sustantivos, adjetivos,…), permitiendo determinar qué tipos de categorías gramaticales predominan o no en el texto. Este resultado está disponible en: "Morfología → Gráficas → Categorías gramaticales".

***III. Flexiones verbales***

Establece cómo se distribuyen las palabras del texto según su flexión verbal (infinitivo, gerundio, presente de indicativo,…), permitiendo reconocer los tiempos verbales más y menos utilizados en el texto. Este resultado está disponible en: "Morfología → Gráficas → Flexiones verbales".

***IV. Flexiones no verbales***

Indica cómo se distribuyen las palabras del texto según su flexión no verbal (palabras en masculino, femenino, singular, plural,…), permitiendo observar el género y número, más y menos frecuentes en el texto. Este resultado está disponible en: "Morfología → Gráficas → Flexiones no verbales".

***V. Distribución de las oraciones y párrafos según la cantidad de verbos***

Proporciona el número de oraciones o párrafos del texto que contienen N verbos. Este resultado está disponible en: "Morfología → Categoría gramatical → Verbos".

***VI. Distribución de las oraciones y párrafos según la cantidad de sustantivos***

Proporciona el número de oraciones o párrafos del texto que contienen N sustantivos. Este resultado está disponible en: "Morfología → Categoría gramatical → Sustantivos".

***VII. Distribución de las oraciones y párrafos según la cantidad de adjetivos***

Proporciona el número de oraciones o párrafos del texto que contienen N adjetivos. Este resultado está disponible en: "Morfología → Categoría gramatical → Adjetivos".

***VIII. Distribución de las oraciones y párrafos según la cantidad de adverbios***

Proporciona el número de oraciones o párrafos del texto que contienen N adverbios. Este resultado está disponible en: "Morfología → Categoría gramatical → Adverbios".

***IX. Distribución de las oraciones y párrafos según la cantidad de pronombres***

Proporciona el número de oraciones o párrafos del texto que contienen N pronombres. Este resultado está disponible en: "Morfología → Categoría gramatical → Pronombres".

***X. Distribución de las oraciones y párrafos según la cantidad de preposiciones***

Proporciona el número de oraciones o párrafos del texto que contienen N preposiciones. Este resultado está disponible en: "Morfología → Categoría gramatical → Preposiciones".

***XI. Distribución de las oraciones y párrafos según la cantidad de artículos***

Proporciona el número de oraciones o párrafos del texto que contienen N artículos. Este resultado está disponible en: "Morfología → Categoría gramatical → Artículos".

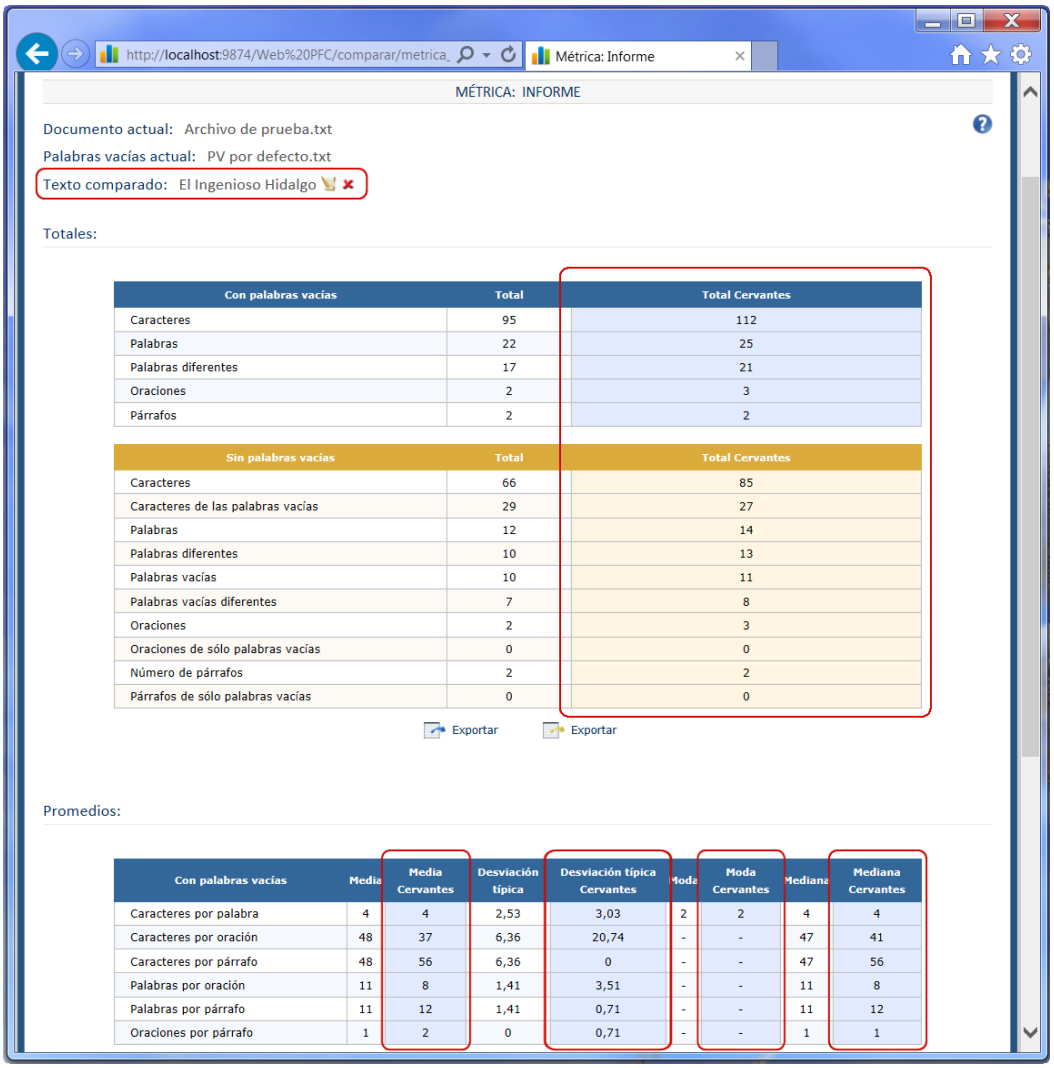
***XII. Distribución de las oraciones y párrafos según la cantidad de conjunciones***

Proporciona el número de oraciones o párrafos del texto que contienen N conjunciones. Este resultado está disponible en: "Morfología → Categoría gramatical → Conjunciones".

**Comparativa (v.2014)**

Esta herramienta proporciona una gran funcionalidad adicional a Paramtext, permitiendo al usuario que compare sus obras con otras de renombre. Su objetivo es ayudar a comprender y valorar los resultados propios mediante la comparación con aquellos que aparecen cuando se parametrizan obras de gran importancia.

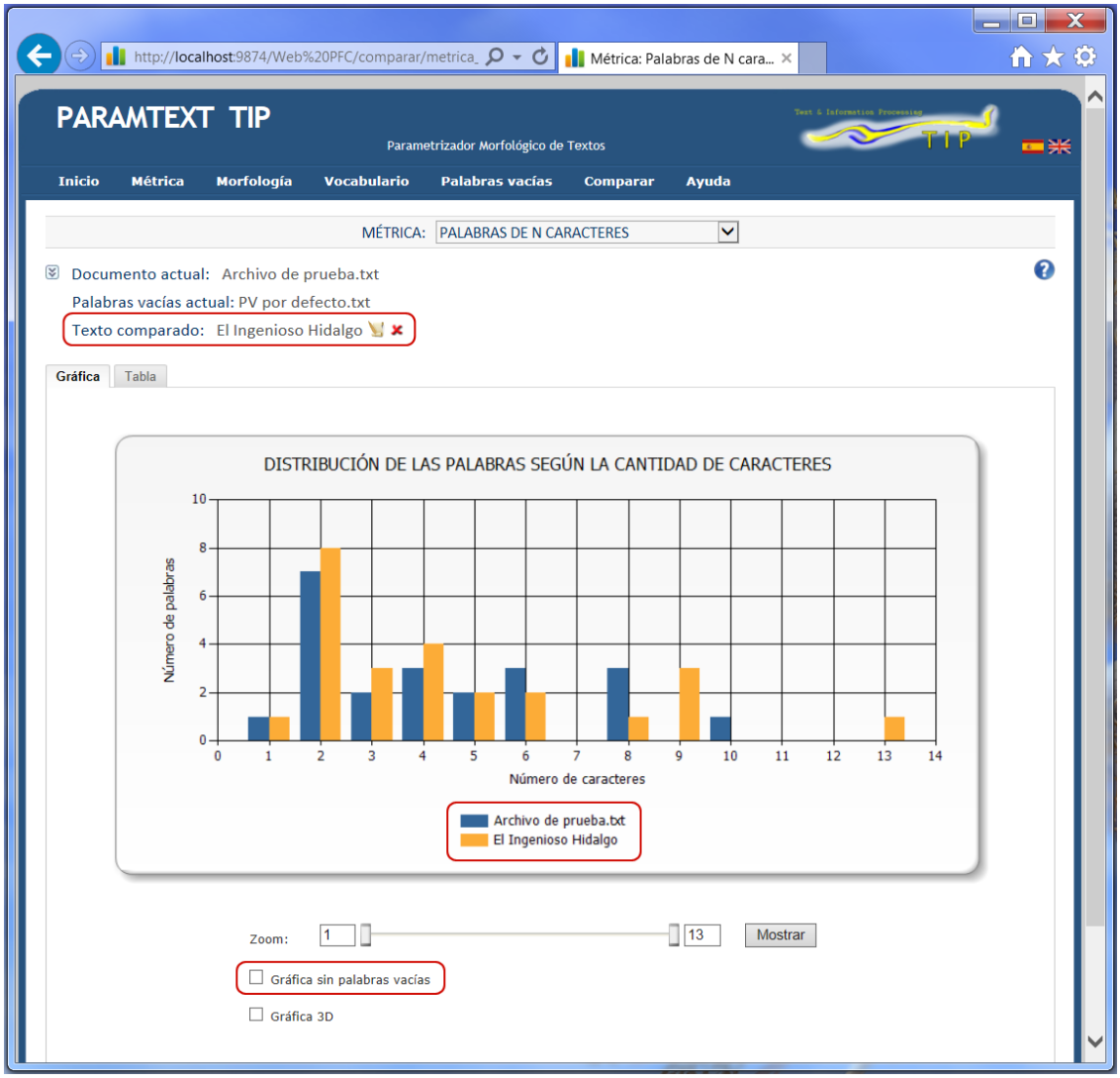
Tras hacer uso de esta herramienta podrán verse en todo momento el nombre de la obra comparada y las mismas tablas y gráficas que en el programa original pero en ellas se mostrarán de forma clara tanto las estadísticas propias como las del texto elegido para realizar la comparación, así como opciones adicionales en caso de ser relevante, como la casilla que permite visualizar la gráfica con o sin palabras vacías en la sección de métrica.




| Con palabras vacías | Total | Total Cervantes |
|---|---|---|
| Caracteres | 95 | 112 |
| Palabras | 22 | 25 |
| Palabras diferentes | 17 | 21 |
| Oraciones | 2 | 3 |
| Párrafos | 2 | 2 |

| Sin palabras vacías | Total | Total Cervantes |
|---|---|---|
| Caracteres | 66 | 85 |
| Caracteres de las palabras vacías | 29 | 27 |
| Palabras | 12 | 14 |
| Palabras diferentes | 10 | 13 |
| Palabras vacías | 10 | 11 |
| Palabras vacías diferentes | 7 | 8 |
| Oraciones | 2 | 3 |
| Oraciones de sólo palabras vacías | 0 | 0 |
| Número de párrafos | 2 | 2 |
| Párrafos de sólo palabras vacías | 0 | 0 |

| Con palabras vacías | Media | Media Cervantes | Desviación típica | Desviación típica Cervantes | Moda | Moda Cervantes | Mediana | Mediana Cervantes |
|---|---|---|---|---|---|---|---|---|
| Caracteres por palabra | 4 | 4 | 2,53 | 3,03 | 2 | 2 | 4 | 4 |
| Caracteres por oración | 48 | 37 | 6,36 | 20,74 | - | - | 47 | 41 |
| Caracteres por párrafo | 48 | 56 | 6,36 | 0 | - | - | 47 | 56 |
| Palabras por oración | 11 | 8 | 1,41 | 3,51 | - | - | 11 | 8 |
| Palabras por párrafo | 11 | 12 | 1,41 | 0,71 | - | - | 11 | 12 |
| Oraciones por párrafo | 1 | 2 | 0 | 0,71 | - | - | 1 | 1 |




### 17.1.9 POLÍTICA DE PRIVACIDAD

ParamText TIP garantiza que los archivos recibidos:

- No serán examinados o modificados en ningún momento.
- No serán facilitados a terceras partes.
- Serán eliminados de nuestros servidores tras un período de inactividad del usuario.