Tesseract para el euskera

Manual

Autor:

www.eleka.net

Índice

1 Introducción

Este documento es el manual de la herramienta para el euskera *Tesseract.*

Esta herramienta es un paquete que permite reconocer textos en euskera escaneados por OCR. Para esta labor, se utiliza la infraestructura patrocinada por Google: la aplicación de OCR Tesseract. Para más información sobre este proyecto puede visitar su página web:

http://code.google.com/p/tesseract-ocr

El objetivo principal es poner en manos de la sociedad una herramienta para reconocer textos escritos en euskera de un modo fiable y automático.

2 Requisitos

2.1 Sistema operativo

En entorno Windows: Windows XP, Windows Vista o Window 7.

En entorno Linux: se ha desarrollado y probado en distribuciones Debian/Ubuntu, pero funciona sin problemas en otras distribuciones.

2.2 Software necesario

En entorno Windows no requiere ningún otro software. Para Linux, consultar el apartado En entorno Linux para ver los pasos de instalación.

2.3 Incompatibilidades

No se ha encontrado incompatibilidad con ningún programa.

3 Instalación

3.1 En entorno Windows

Para instalar el *OCR de código abierto para el* euskera haga doble clic en el instalador. La instalación es automática. El proceso de instalación nos llevará a través de las pantallas que aparecen a continuación.

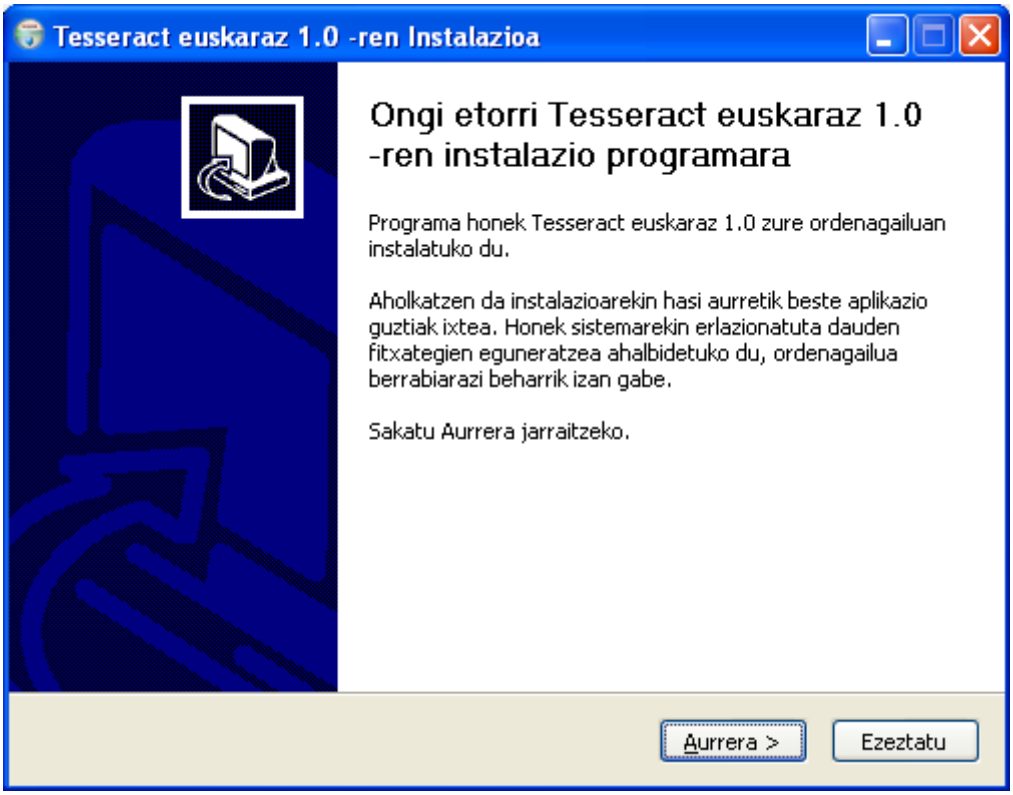

Imagen 1: Ventana de bienvenida del instalador

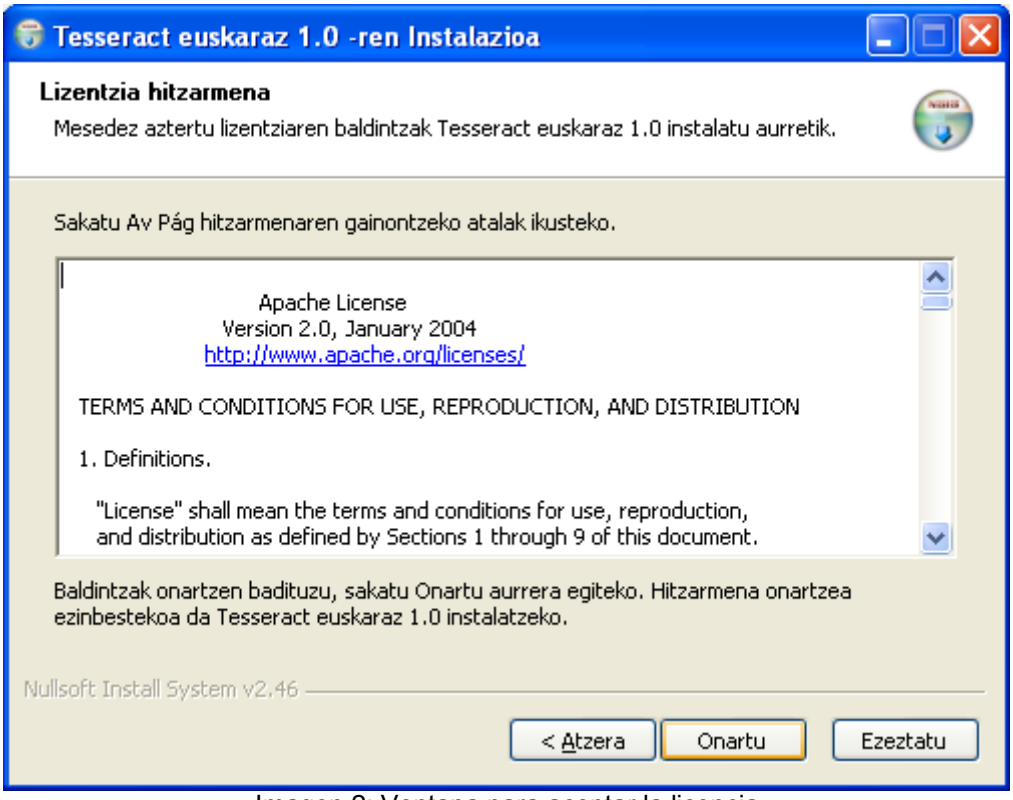

Imagen 2: Ventana para aceptar la licencia

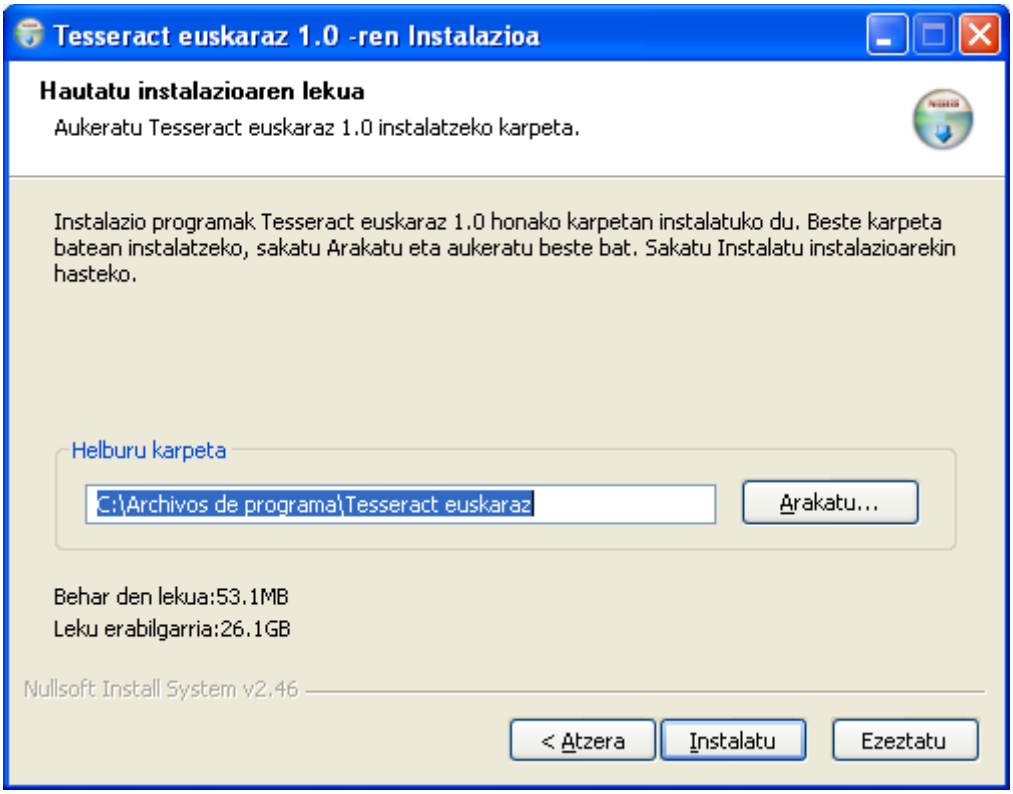

Imagen 3: Ventana para la selección de directorio de instalación

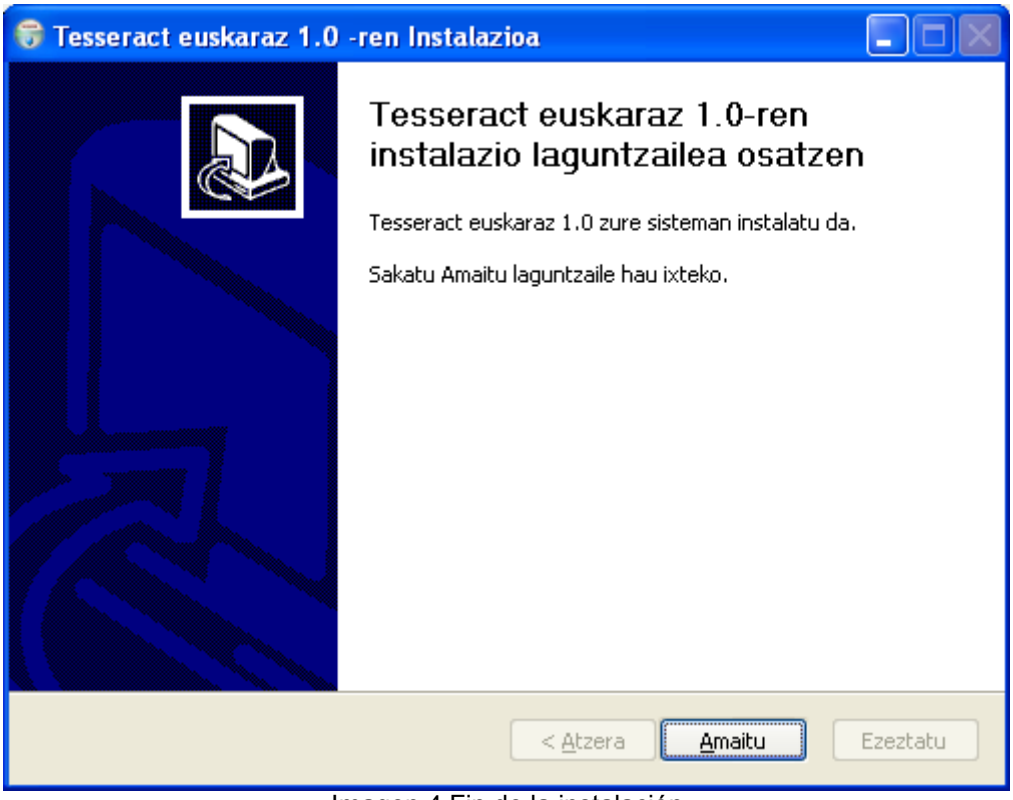

Imagen 4 Fin de la instalación

3.2 En entorno Linux

En entorno Linux será necesario compilar el código fuente. Para ello, tendremos que instalar ciertas herramientas de desarrollo, en caso de que no dispongamos ya de ellas:

```
$ sudo apt-get install build-essential
```

3.2.1 Instalaciones previas necesarias

Leptonica

Se trata de una librería de código abierto que dispone de algoritmos para el procesamiento y análisis de imágenes (http://www.leptonica.org). Para instalarla, tendremos que descargar el código fuente y compilarlo. Habrá que instalar previamente las siguientes librerías:

```
$ sudo apt-get install libpng12-dev libjpeg62-dev libtiff4-dev
  libgif-dev zlib1g-dev
```

Para descargar el código fuente, ejecutar el siguiente comando:

```
$ wget http://www.leptonica.org/source/leptonica-1.68.tar.gz
```

Una vez se haya descargado el código fuente, debemos descomprimirlo, compilarlo e instalarlo. Para ello seguiremos los siguientes pasos:

```
$ tar -zxvf leptonica-1.68.tar.gz
$ cd  leptonica-1.68
$ ./configure
$ make
$ sudo make install
$ sudo ldconfig
```

3.2.2 Tesseract

En primer lugar, descargar el código fuente:

```
$ wget http://tesseract-ocr.googlecode.com/files/tesseract-3.00.tar.gz
```

Después, lo descomprimiremos, lo compilaremos e instalaremos:

```
$ tar -zxvf  tesseract-3.00.tar.gz
$ cd   tesseract-3.00
$ ./runautoconf
$ ./configure
$ make
$ sudo make install
$ sudo ldconfig
```

Una vez hayamos instalado Tesseract, deberemos descargar los archivos de datos para el euskera desde la página web de HPS. Después continuar con los siguientes pasos:

```
$ sudo cp eus.traineddata /usr/local/share/tessdata
```

Puede ser que en la instalación de Tesseract nos falte un archivo llamado *hocr*, el cual debería estar en la carpeta /usr/local/share/tessdata/configs. Si es así, podemos crearlo nosotros mismos. Para ello, deberemos abrir nuestro editor de textos favorito y escribir la siguiente línea:

```
tessedit_create_hocr 1
```

Después, lo guardaremos en el archivo /usr/local/share/tessdata/configs/hocr.

3.2.3 tesseractgui

La interfaz de la aplicación se ejecuta dentro XULRunner, un paquete de ejecución de la Fundación Mozzila. Para instalarlo debemos ejecutar el

siguiente comando:

```
$ sudo apt-get install xulrunner-2.0
```

Desde la página web de HPS, descargaremos los archivos de la interfaz de la aplicación, y lo descomprimiremos:

```
$ tar -zxvf tesseractgui.tar.gz
```

Para iniciar la aplicación:

```
$ cd tesseractgui
$ xulrunner application.ini
```

4 Desinstalación

4.1 En entorno Windows

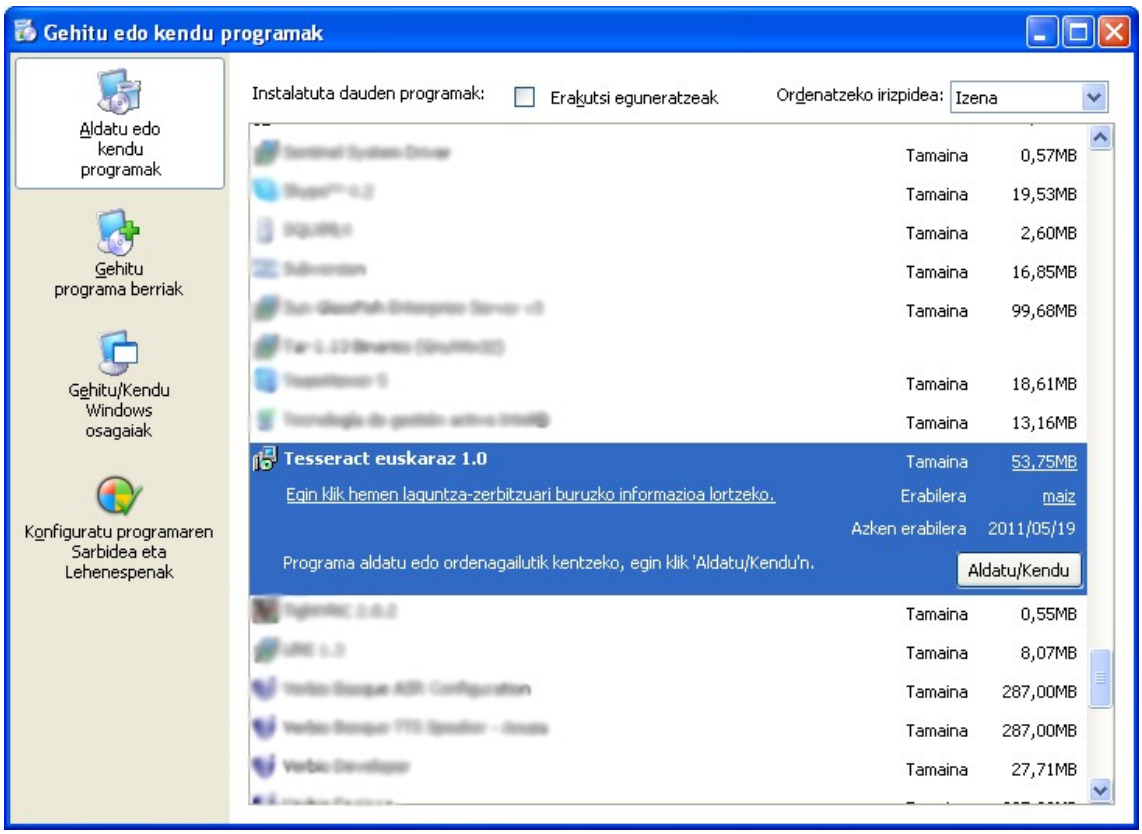

Imagen 5: Desinstalando el OCR

El proceso para desinstalar el OCR de código abierto para el euskera es el mismo que se utiliza para desinstalar cualquier otra aplicación en Windows: basta con ir al “panel de control”, “agregar o quitar programas”, seleccionar *Tesseract euskaraz* y seguir los pasos del desinstalador.

4.2 En entorno Linux

Si el código fuente está en el sistema, bastará con ejecutar el siguiente comando en la línea de comando para cada proyecto, desde la carpeta de cada proyecto:

```
$ sudo make uninstall
```

En el caso de tesseractgui deberemos eliminarlo directamente desde la línea

de comando:

```
$ sudo rm tesseractgui
```

Si el código fuente no está en el sistema, debemos borrar manualmente los ficheros de instalación:

```
$ sudo rm -Rf /usr/local/include/tesseract
$ sudo rm -Rf /usr/local/include/leptonica
$ sudo rm -Rf /usr/local/share/tessdata
$ sudo rm /usr/local/lib/liblept.a
$ sudo rm /usr/local/lib/libtesseract_*
$ sudo rm /usr/local/bin/cntraining
$ sudo rm /usr/local/bin/mftraining
$ sudo rm /usr/local/bin/unicharset_extractor
$ sudo rm /usr/local/bin/wordlist2dawg
$ sudo rm /usr/local/bin/combine_tessdata
$ sudo rm /usr/local/bin/tesseract
```

5 Modo de empleo

5.1 En entornos Windows y Linux

Proceso de OCR

El modo de empleo de *Tesseract para el euskera* es bastante simple, una vez lo hayamos instalado.

Para ponerlo en marcha, bastará con hacer doble clic sobre el icono del programa y accederemos a la siguiente pantalla (en entorno Linux abrir un terminal y ejecutar *xulrunner application.ini*):

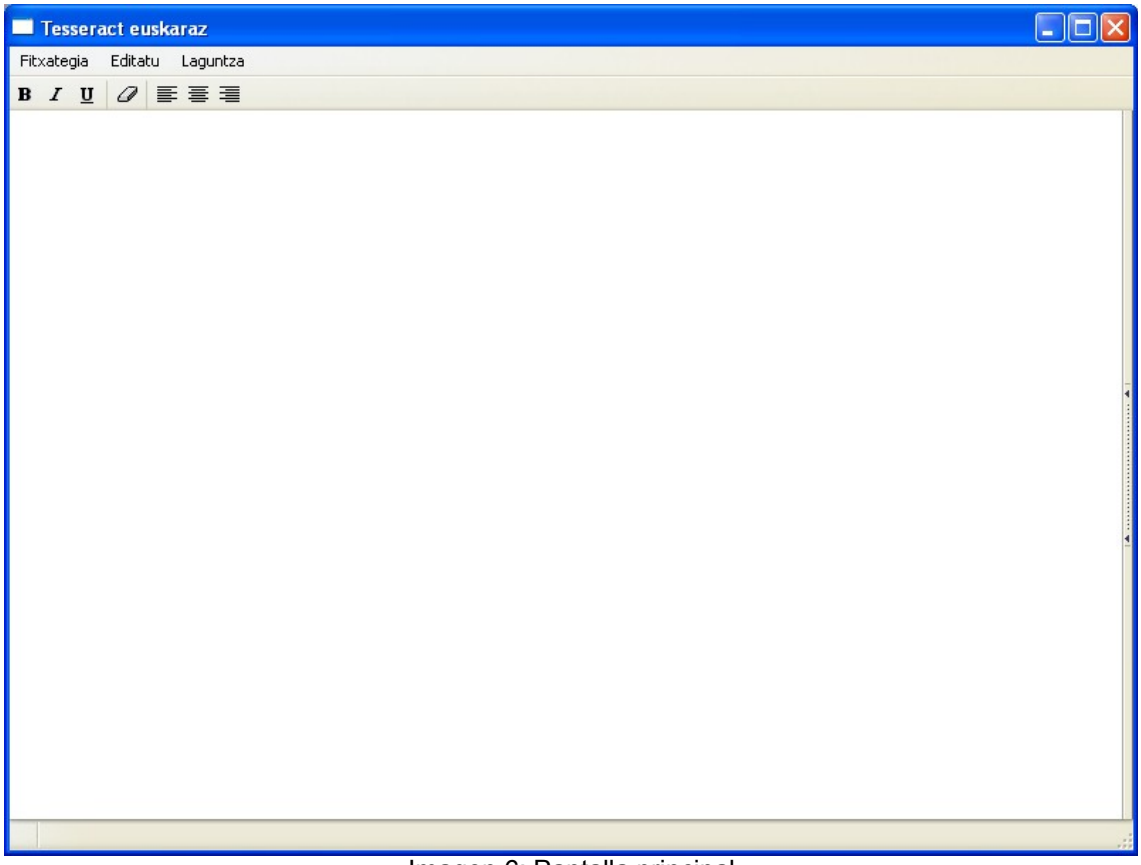

Imagen 6: Pantalla principal

Para poner en marcha el proceso de OCR, seleccionamos la opción *Abrir* del menú *Archivo*, y seleccionamos una imagen. Los formatos de imagen permitidos son JPG, GIF, PNG o TIFF[1]. Una vez seleccionada la imagen, se pondrá en marcha el proceso de OCR. El resultado se mostrará en la parte

1El formato *TIFF* está permitido para el proceso de OCR pero no es posible mostrarlo en pantalla.

derecha de la pantalla, y en la izquierda tendremos la imagen que hemos pasado por el OCR.

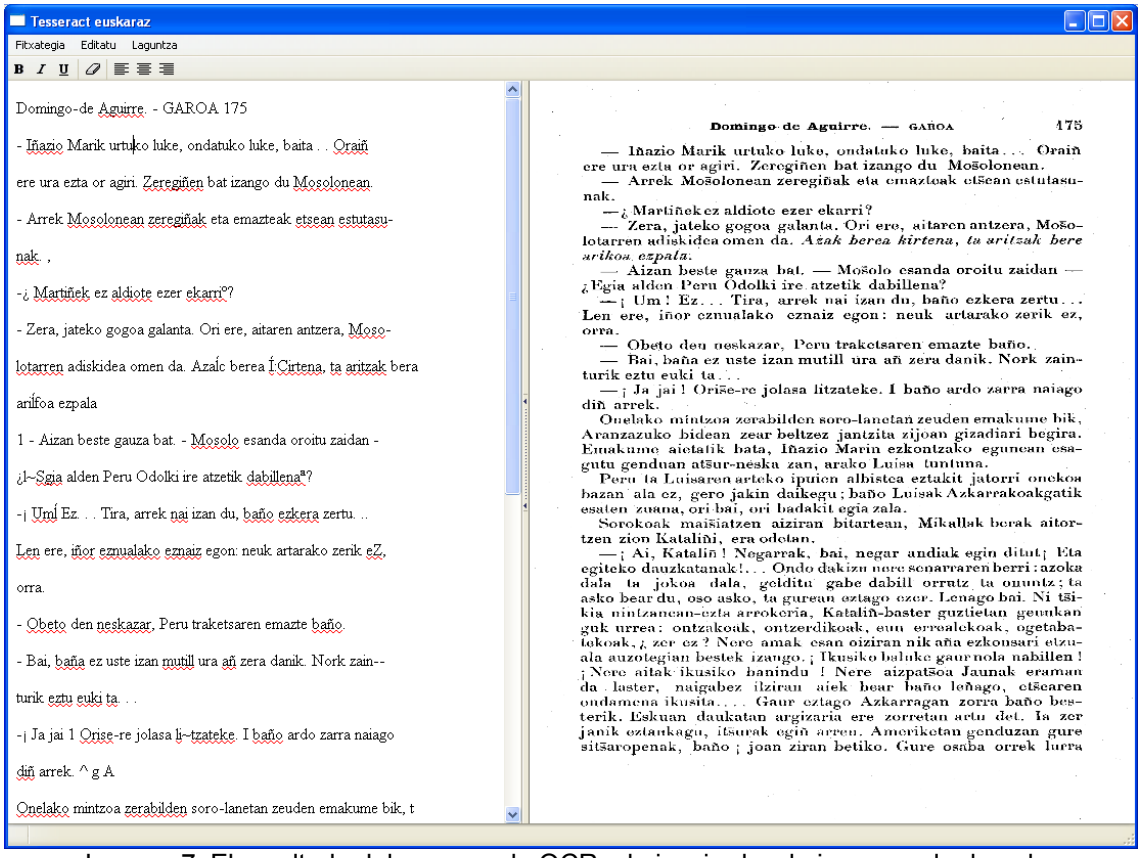

Imagen 7: El resultado del proceso de OCR a la izquierda y la imagen a la derecha

El resultado es editable, y podemos actualizarlo/corregirlo desde la propia interfaz. Como herramienta complementaria se ofrece un corrector ortográfico.

Una vez hayamos terminado la edición, podremos guardar el resultado en un archivo HTML, utilizando para ello la opción *Guardar* del menú *Archivo*.

Opciones

La aplicación dispone de algunas opciones de personalización, accesibles desde el menú *Edición* > *Opciones*.

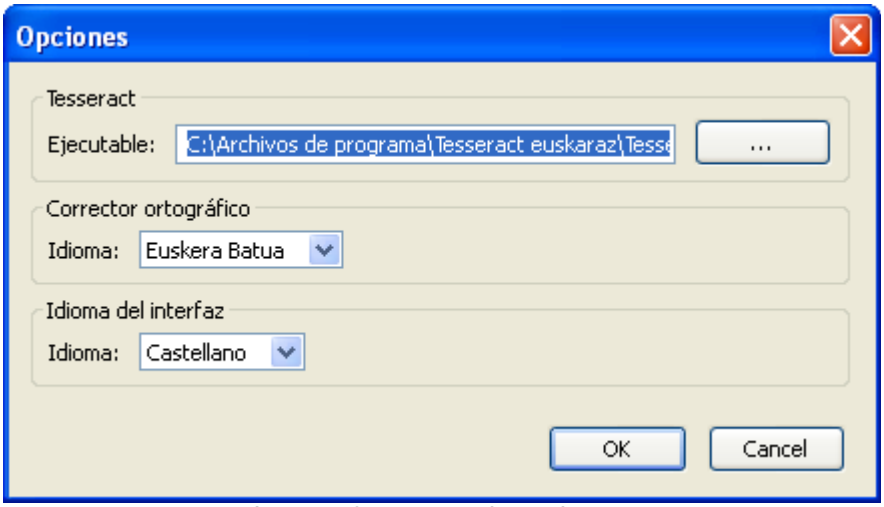

Imagen 8: ventana de opciones

Desde esta pantalla podemos actualizar las siguientes opciones:

- Ubicación de Tesseract: para el proceso de OCR se utiliza la aplicación *Tesseract*. Desde esta opción definiremos su ubicación.
- Corrector ortográfico: la aplicación dispone de 2 correctores ortográficos, una para el Euskera Batua y otro para el Vizcaíno. Desde aquí seleccionaremos el que queramos utilizar.
- Idioma de la interfaz: la aplicación se puede mostrar tanto en euskera como en castellano.