

Tesseract euskaraz

Eskuliburua



Egilea:

Aurkibidea

1	Hitzaurrea.....	3
2	Eskakizunak.....	4
2.1	Sistema eragilea.....	4
2.2	Beharrezko softwarea.....	4
2.3	Bateraezintasunak.....	4
3	Instalazioa.....	5
3.1	Windows ingurunean.....	5
3.2	Linux ingurunean.....	7
3.2.1	Aurretik instalatu beharrekoak.....	7
	Leptonica.....	7
3.2.2	Tesseract.....	8
3.2.3	tesseractgui.....	9
4	Desinstalazioa.....	10
4.1	Windows ingurunean.....	10
4.2	Linux ingurunean.....	10
5	Erabilera.....	12
5.1	Windows nahiz Linux ingurunean.....	12
	OCR prozesua.....	12
	Aukerak.....	13

1 Hitzaurrea

Dokumentu hau *Tesseract euskaraz* tresnaren eskuliburua da.

Tresna hau euskarazko eskaneatutako testuak OCR bidez ezagutzeko gai den pakete bat da. Horretarako erabili den azpiegitura Googlek babesturiko Tesseract OCR tresna izan da. Tresna honen inguruko informazio gehiago hemen aurkitu daiteke:

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

Helburu nagusia euskaraz idatzitako testuak modu fidagarri eta automatikoan ezagutzeko gai izango den tresna gizartearen eskuetan jartzea izan da.

2 Eskakizunak

2.1 Sistema eragilea

Windows ingurunean: Windows XP, Windows Vista edo Windows 7.

Linux ingurunean: Debian/Ubuntu banaketetan garatu eta probatu da, baina bestelakoekin funtzionatzeko arazorik ez dago.

2.2 Beharrezko softwarea

Windows ingurunean ez da aparteko softwarerik behar. Linux ingurunearen kasuan ikusi instalaziorako pausuak Linux ingurunean atalean.

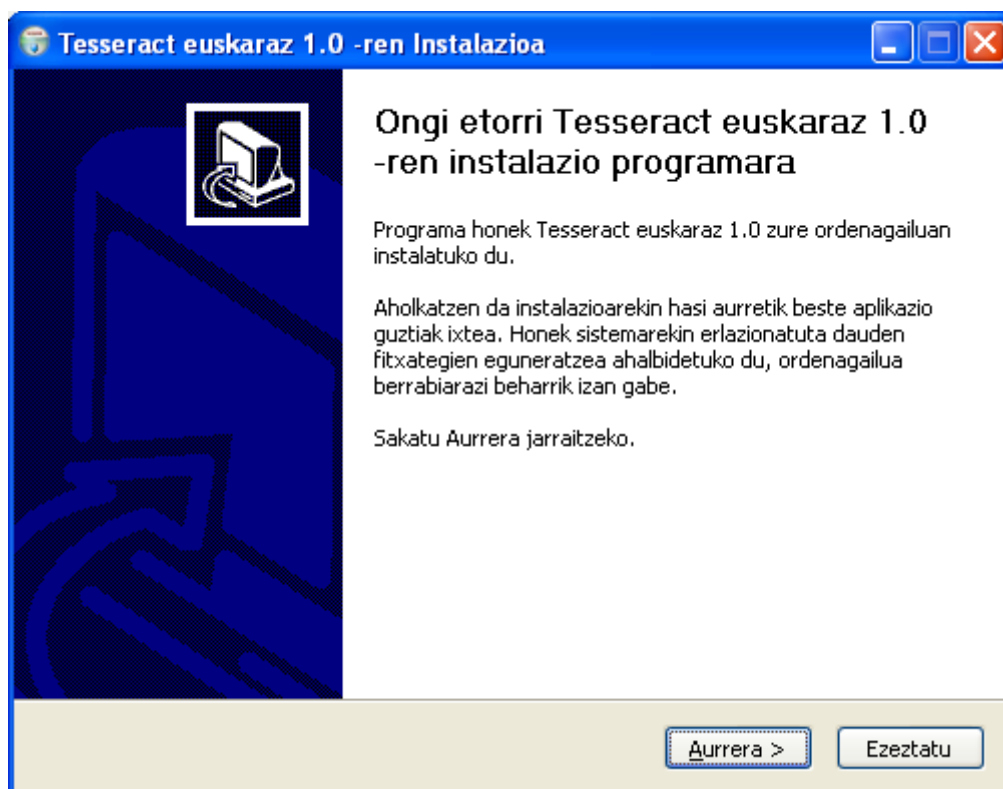
2.3 Bateriaezintasunak

Ez da inongo programarekin bateraezintasunik topatu.

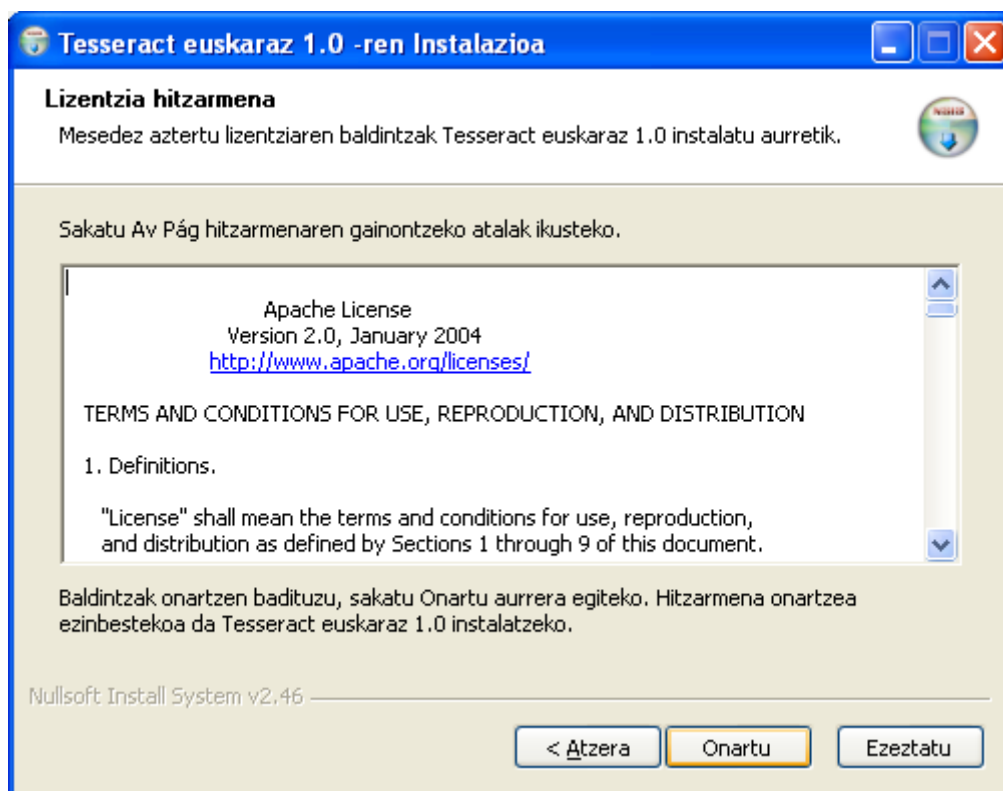
3 Instalazioa

3.1 Windows ingurunean

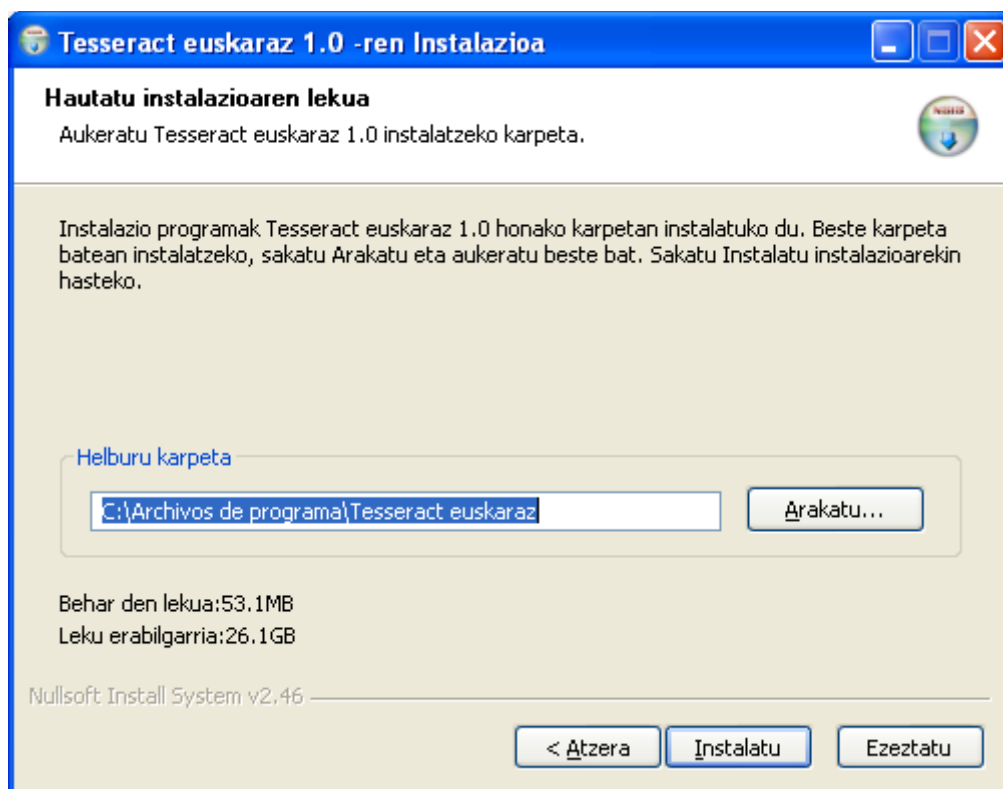
Kode irekiko euskarazko OCRa instalatzeko klik bikoitza egin instalatzailean. Instalazioa automatikoa da. Instalazio prozesuak ondorengo irudietan agertzen diren pantailetan zehar eramango gaitu.



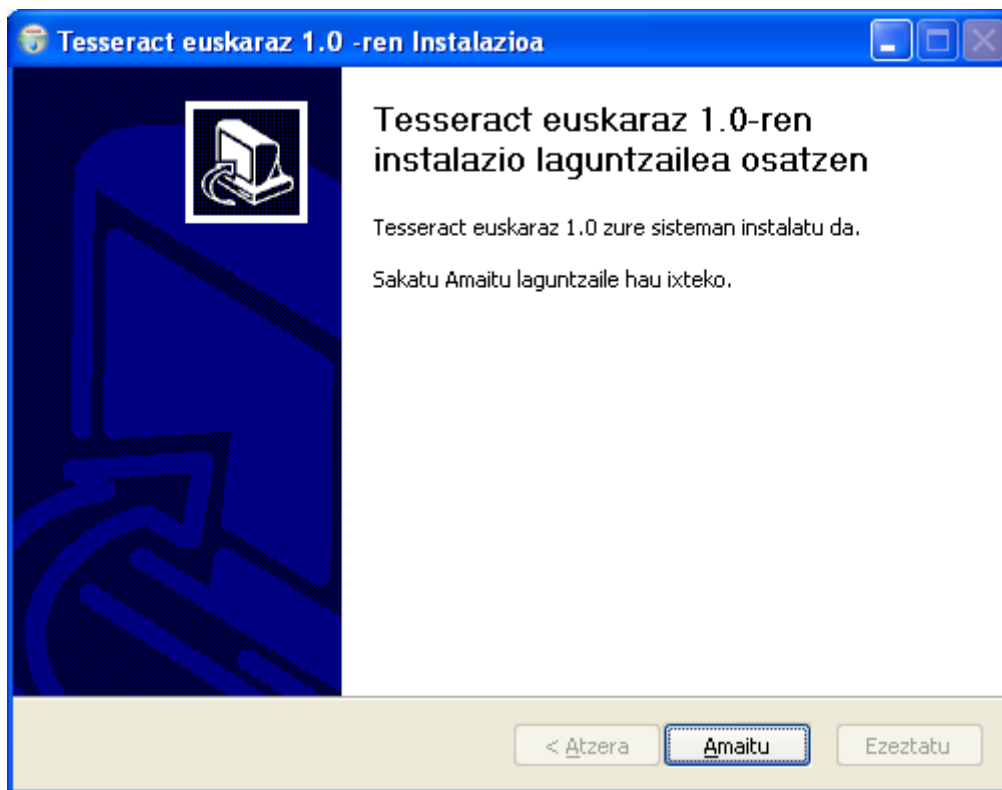
1. Irudia: Instalatzailearen ongi-etorri leihoa



2. Irudia: Lizentziaren onarpen leihoa



3. Irudia: Direktorioaren aukeraketa leihoa



4. Irudia: Instalazioaren amaiera

3.2 Linux ingurunean

Linux ingurunean instalatzeko beharrezkoa da iturburu kodea konpilatzea. Horretarako lehenik eta behin garapenerako erremintak instalatu beharko dira lehenagotik instalaturik ez bazeuden:

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

3.2.1 Aurretik instalatu beharrekoak

Leptonica

Irudien prozesamendurako eta analisirako kode irekiko liburutegia da (<http://www.leptonica.org>). Instalatzeko iturburu kodea jaitsi eta konpilatu behar da. Aldez aurretik ondoko liburutegiak instalatu behar dira:

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

Iturburu kodea jaisteko ondoko agindua exekutatu:

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

Behin iturburu kodea jaitsi dela, erauzi, konpilatu eta instalatu daiteke:

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

3.2.2 Tesseract

Lehenik eta behin iturburu kodea jaitsi:

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

Ondoren, jaitsi berri den fitxategia erauzi, karpetan sartu eta konpilatu eta instalatu daiteke:

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

Behin Tesseract instalatu dela euskararentzako datu-fitxategiak deskargatu beharko dira HPSren webgunetik. Ondoren jarraian datozen pausuak exekutatu:

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

Baliteke Tesseract-en *hocr* izeneko fitxategi bat falta izatea, /usr/local/share/tessdata/configs direktorioan egon beharko luke. Horrela bada gu geuk sortu dezakegu. Gure gogoko testu editorea ireki, eta ondoko lerroa idatzi:

```
tessedit_create_hocr 1
```

Ondoren gorde esandako fitxategian:

```
/usr/local/share/tessdata/configs/hocr
```


3.2.3 tesseractgui

Aplikazioaren interfazeak Mozilla Fundazioaren XULRunner exekuzio ingurunea erabiltzen du. Horregatik, lehenik eta behin, hau instalatu beharko dugu:

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

Ondoren, HPSren webgunetik, interfazeari dagokion fitxategia jaitsi eta ondoko pausuak jarraitu:

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

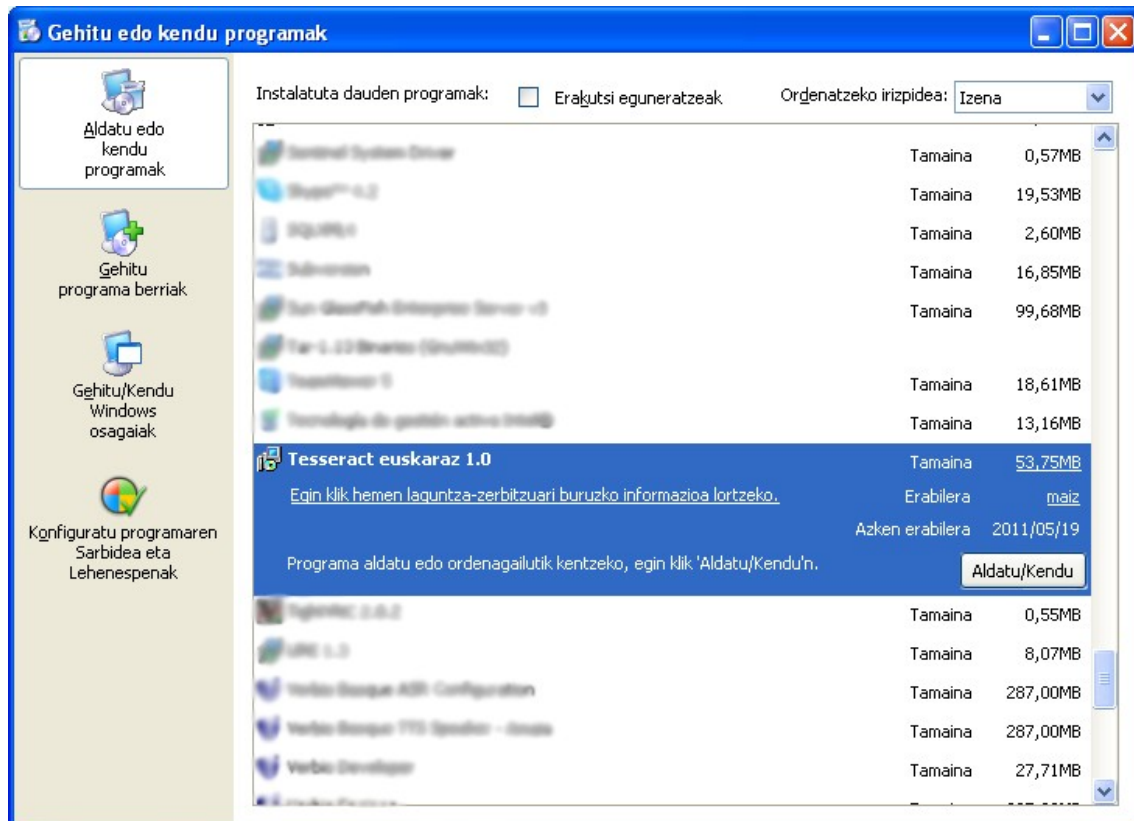
Aplikazioa martxan jartzeko ondoko agindua exekutatu:

```
$ cd tesseractgui
```

```
$ xulrunner application.ini
```

4 Desinstalazioa

4.1 Windows ingurunean



5. Irudia: *Tesseract euskaraz* desinstalatu

Kode irekiko euskarazko OCRa desinstalatzeko prozedura edozein Windows aplikazioa desinstalatzeko jarraitu beharreko berbera da: nahikoa da “kontrol-panela”eko “Gehitu edo kendu programak” leihoan *Tesseract euskaraz* aukeratu eta desinstalatzailearen pausoak jarraitzea.

4.2 Linux ingurunean

Iturburu koda sisteman badago, nahikoa da proiektu bakoitzeko ondoko agindua exekutatzea komando lerrotik, proiektu bakoitzaren direktorioaren barruan gaudela:

```
$ sudo make uninstall
```

Interfazearen kasuan, bere karpeta ezabatzearekin nahikoa da:

```
$ rm -Rf tesseractgui
```

Iturburu koderik ez badugu, banan-banan ezabatuko ditugu instalazio

fitxategiak:

```
$ 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
```

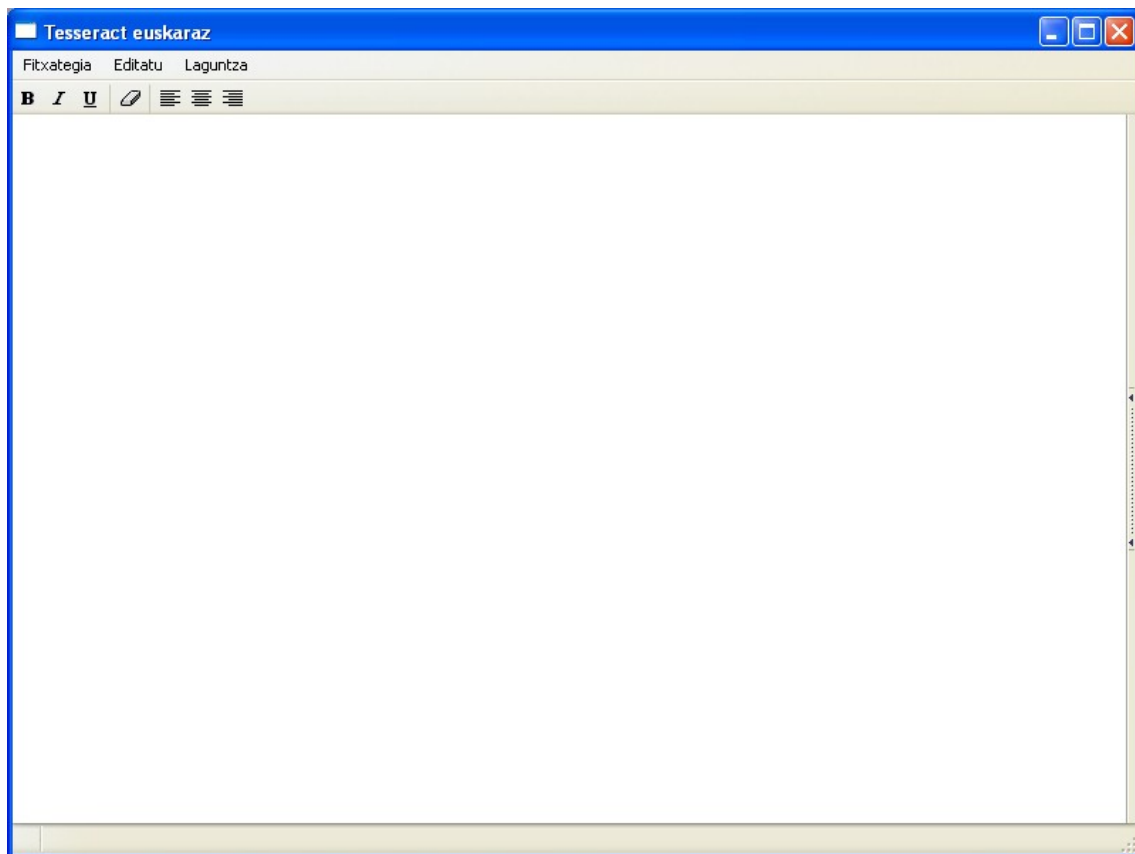
5 Erabilera

5.1 Windows nahiz Linux ingurunean

OCR prozesua

Behin *Tesseract euskaraz* instalatuta dugularik, bere erabilera nahiko sinplea da.

Programa martxan jartzeko klik bikoitza egin behar da programaren ikonoan eta honako pantaila ikusiko dugu (Linux ingurunearen kasuan terminal bat zabaltu eta *xulrunner application.ini* agindua exekutatu):

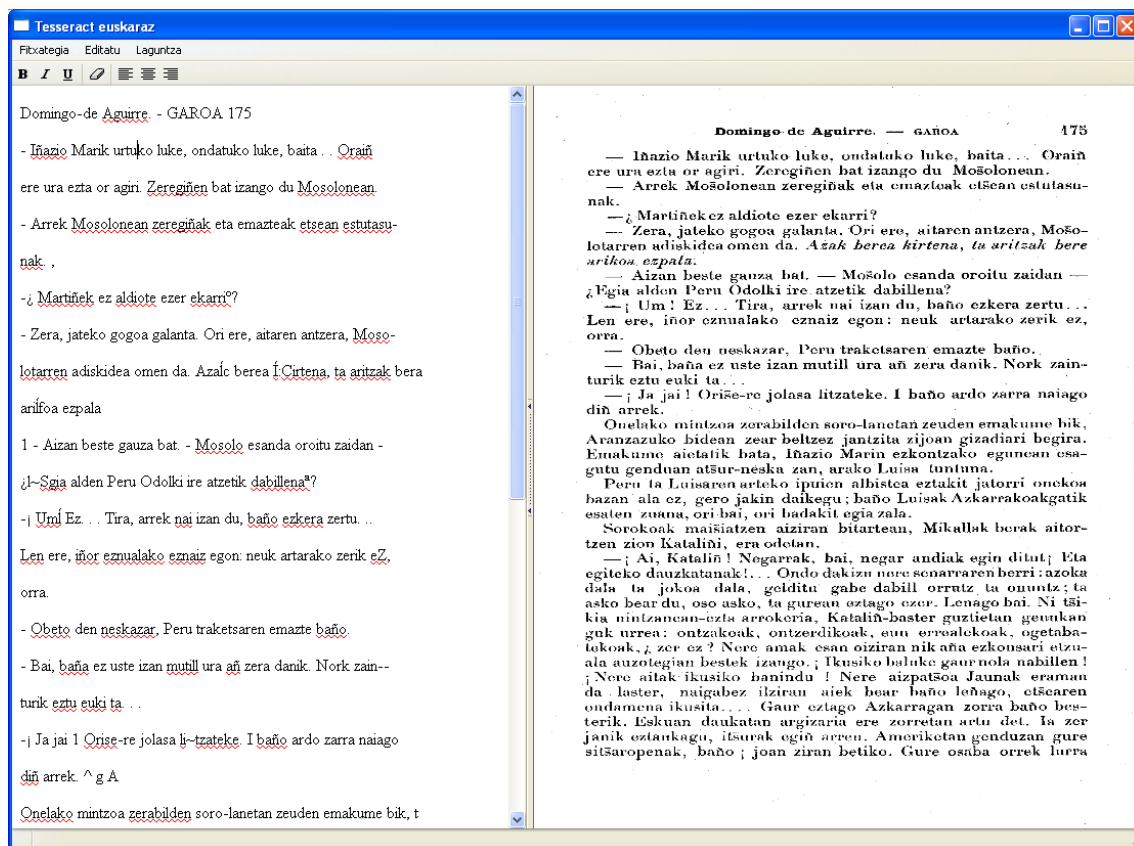


6. Irudia: Pantaila nagusia

OCR prozesua martxan jartzeko, *Fitxategia* menu ireki, *Ireki* aukera sakatu eta irudi bat aukeratzeko leihoan, JPG, GIF, PNG edo TIFF¹ motako irudi bat aukeratu. Irudia aukeratzearekin batera prozesua martxan jarriko da eta emaitza pantaila nagusiaren ezkerrean aurkeztuko zaigu. Emaitzarekin batera, pantaila nagusiaren eskuinaldean aukeratu dugun irudia erakutsiko

¹TIFF motako irudiak OCR prozesutik pasa daitezke, baina ezin dira pantailan erakutsi.

zaigu.



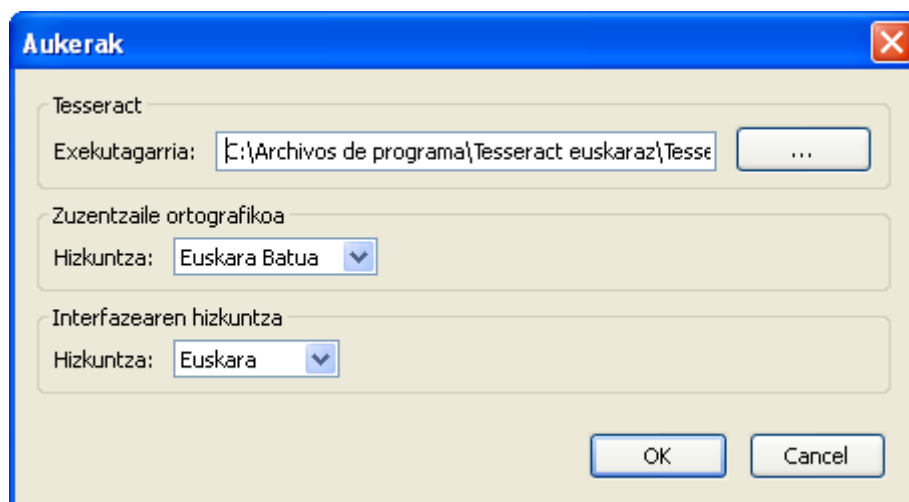
7. Irudia: OCR prozesuaren emaitza ezker aldean eta irudia eskuinaldean

Emaitza editagarria da eta interfaze honetatik bertatik editatu ahal izango dugu. Erreminta baliagarri bezala zuzentzaile ortografiko bat eskaintzen zaigu.

Behin emaitza prest dugula, *Fitxategia* > *Gorde* aukera sakatuta, HTML fitxategi batean gorde ahal izango dugu.

Aukerak

Interfazeak zenbait pertsonalizazio aukera ematen dizkigu. Honetarako, sakatu *Editatu* > *Aukerak* menua.



8. Irudia: Aukerak pantaila

Pantaila honetan aukera hauek eguneratu ditzakegu:

- Tesseract exekutagarriaren kokapena: esan bezala OCR prozesurako erabiltzen den tresna *Tesseract* da. Aukera honetatik bere kokapena zehaztu dezakegu.
- Zuzentzaile ortografikoa: 2 zuzentzaile ortografiko erabiltzeko aukera ematen du aplikazioak, Euskara Baturako edo Bizkaierarako. Honekin zehaztuko dugu zein erabili nahi dugun.
- Interfazearen hizkuntza: Aplikazioa euskaraz nahiz gazteleraz erabili dezakegu.