ingeniaritza ELEKA linguistikoa

Tesseract-OCRopus euskaraz

Eskuliburua

Egilea:

ingeniaritza ELEKA linguistikoa

www.eleka.net

Aurkibidea

1 Hitzaurrea

Dokumentu hau *Tesseract-OCRopus euskaraz* tresnaren eskuliburua da.

Tresna hau euskarazko eskaneatutako testuak OCR bidez ezagutzeko gai izango den pakete bat da. Horretarako erabiliko den azpiegitura Googlek babesturiko Tesseract oinarrizko OCR tresna eta OCRopus dokumentu analizatzailea izango dira.

Tesseract eta Ocropusen arteko harremana azaltzea argigarria izango da puntu honetan. Horretarako, egokiena bakoitzaren funtzio eta betebeharrak azaltzea izango da:

- **Tesseract**: OCR motor soila da. Teknologia aurreratuenetan oinarritzen da eta oso fidagarria da. Bere lana testu zati bat hartu eta testu horretako esaldiak ezagutzea da. Tesseractek testu sinplea bakarrik ulertzen du, hau da, zutabe bakarrean eta argazkirik gabe aurkezten den testua.

- **OCRopus**: dokumentu konplexuak analizatzeko tresna. Dokumentuak analizatzeak hiru pauso eskatzen ditu oro har:

1. Lehenengo eta behin, dokumentua osatzen duten zutabeak, orri buruak, orri-oinak, irudiak, taulak, etab. identifikatu behar dira eta beraien arteko erlazioak zehaztu. Erlazio horiek garrantzitsuak dira testuaren fluxuan eragina duten heinean. Izan ere, orri-oina eta orriaren gorputza independenteak dira normalean baina ezkerreko zutabeak eskuinekoaren aurretik irakurri behar dira.

2. Bigarren pausua dokumentua osatzen duten zatiak hartu eta, testua badira, OCR motorrari pasatzea da ezagutu ditzan. Hau da prozesuaren atal konplexuena, Tesseract bidez burutuko dena. Une honetan Tesseract ingelesezko testuak ulertzeko gai da bakarrik. Proiektuaren helburua euskarazko testuekin ingelesekin duen fidagarritasun maila bera lortzea da.

3. Hirugarren eta azken pausoa aurreko bietako informazioa uztartu eta jatorrizko dokumentuaren bertsio elektroniko ahal bezain fidela osatzea da. Ocropusek HTML dokumentu bat (web orri bat) sortzen du eta jatorrizko dokumentuaren estiloa mantentzen saiatzen da.

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

Bi proiektu hauen inguruko informazio gehiago hemen aurkitu daiteke:

- Tesseract proiektuaren webgunea: http://code.google.com/p/tesseract-ocr
- OCRopus proiektuaren webgunea: http://code.google.com/p/ocropus

2 Eskakizunak

2.1 Sistema eragilea

Windows ingurunean: Windows 2000, Windows XP edo Windows Vista.

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.

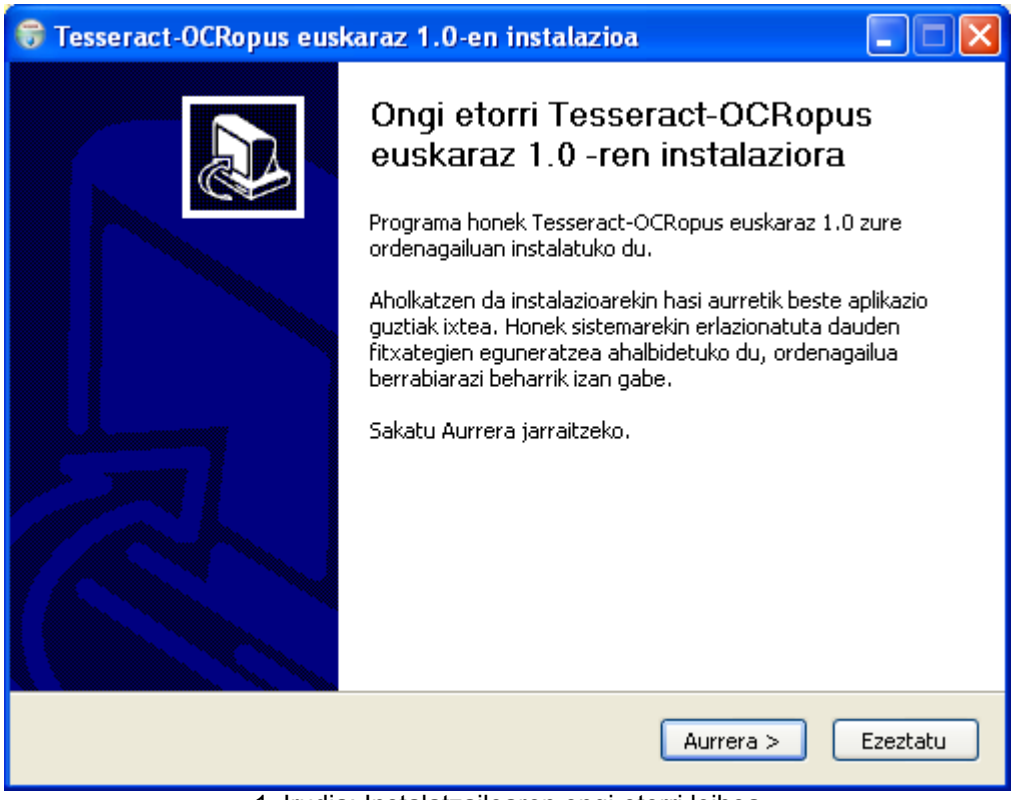
2.3 Bateraezintasunak

Ez da inongo programarekin bateraezintasunik topatu.

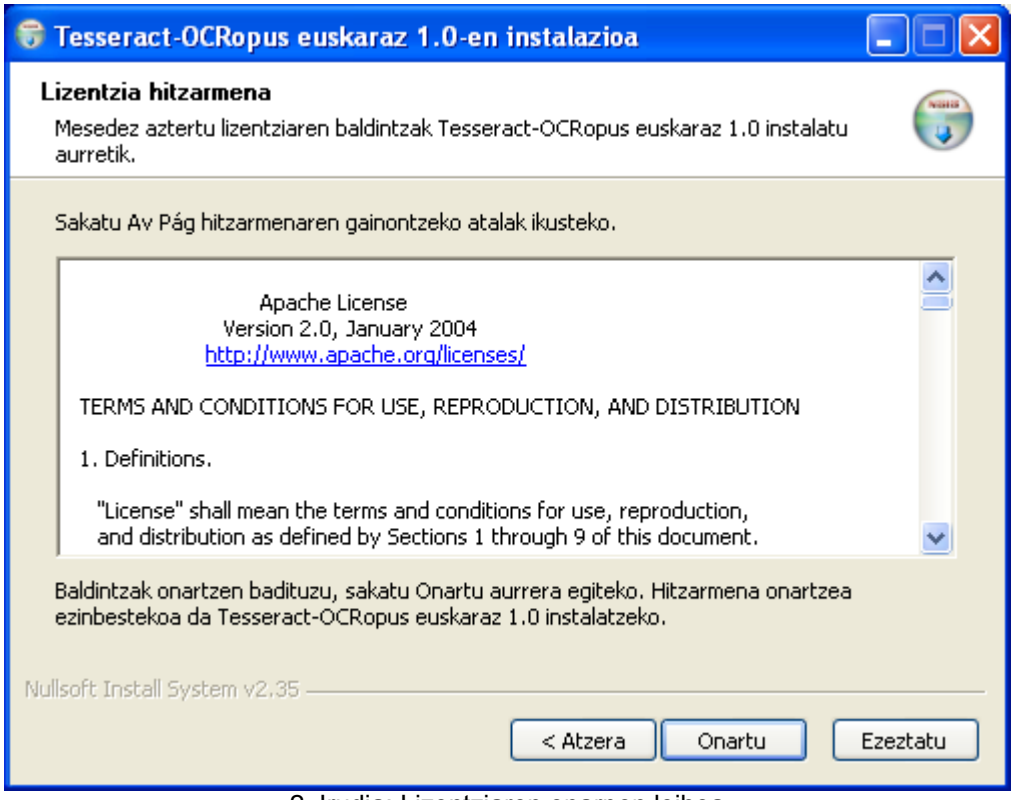
3 Instalazioa

3.1 Windows ingurunean

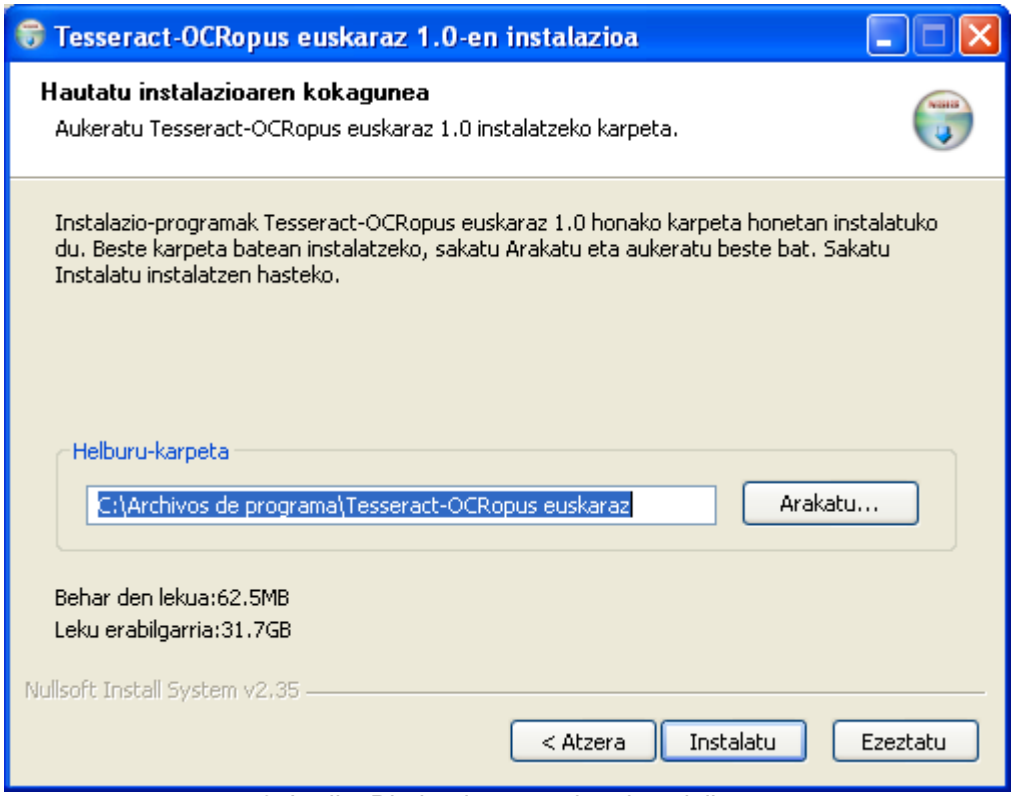
Kode irekiko euskarazko OCRa instalatzeko klik bikoitza egin instalatzailean. Instalazioa automatikoa da. Instalazio prozesuak ondorengo irudietan agertzen diren pantailetan zehar eramango gaitu.



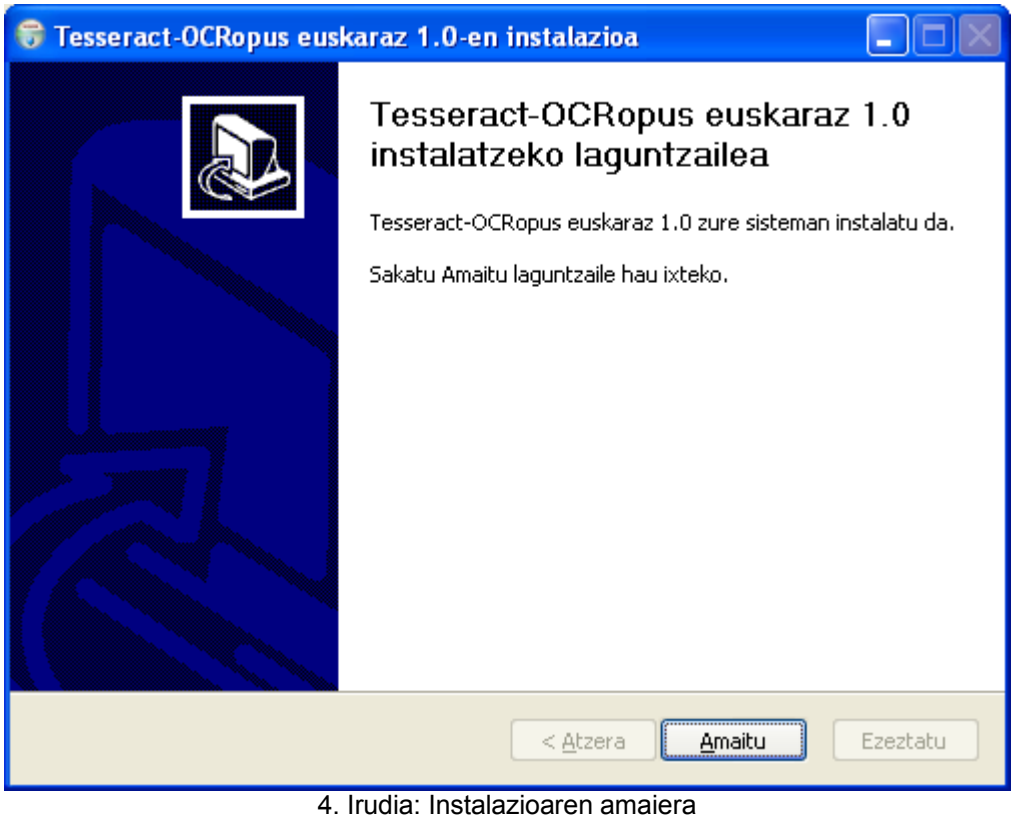
1. Irudia: Instalatzailearen ongi-etorri leihoa



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
$ sudo apt-get install svn
```

3.2.1 AURRETIK INSTALATU BEHARREKOAK

Hunspell

Kode irekiko zuzentzaile ortografikoa da Hunspell (http://hunspell.sourceforge.net). Bi aukera daude Hunspell liburutegia instalatzeko, sistemaren errepositorioak erabilita edo kode iturburua jaitsita. Lehenengo aukera erabilita posible da azken bertsioa ez izatea (1.2.8).

Errepositorioak erabilita:

```
$ sudo apt-get install hunspell-dev
```

Kode iturburua jaitsita:

```
$ wget http://downloads.sourceforge.net/hunspell/hunspell-1.2.8.tar.gz
$ tar -zxvf hunspell-1.2.8.tar.gz
$ cd hunspell-1.2.8
$ ./configure
$ make
$ sudo make install
```

Agian esteka sinboliko bat egin beharko da kode iturburua konpilatu bada:

```
$ sudo ln /usr/local/lib/libhunspell-1.2.so
  /usr/local/lib/libhunspell.so
```

iulib

Irudien ulermenerako algoritmoak dituen kode irekiko liburutegia da (http://code.google.com/p/iulib). 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
  libavcodec-dev libavformat-dev libsdl-gfx1.2-dev libsdl-
  image1.2-dev
```

Iturburu kodea jaisteko ondoko agindua exekutatu:

```
$ svn checkout http://iulib.googlecode.com/svn/trunk/ iulib-read-
  only
```

Behin iturburu kodea jaitsi dela, sortu berri den direktorioan sartu eta konpilatu eta instalatu daiteke:

```
$ cd iulib-read-only
$ ./configure
$ make
$ sudo make install
```

3.2.2 TESSERACT

Lehenik eta behin iturburu kodea jaitsi:

```
$ svn checkout http://tesseract-ocr.googlecode.com/svn/trunk/
  tesseract-ocr-read-only
```

Ondoren sortu berri den karpetan sartu eta konpilatu eta instalatu daiteke:

```
$ cd  tesseract-ocr-read-only
```

Tesseract-ek Hunspell erabiltzea nahi bada (kasu gehienetan emaitza hobeak lortzen dira euskararen kasuan) hiztegi nagusi bezala jarraian azaltzen diren pausoak jarraitu:

```
$ patch -p0 < tess_hunspell.patch
$ aclocal
$ autoheader
$ autoconf
$ automake --add-missing
$ ./configure --with-hunspell LDFLAGS="$LDFLAGS -L/usr/local/lib"
$ make
$ sudo make install
```

Tesseract bere horretan erabili nahi badugu, Hunspell erabili gabe:

```
$ ./configure
$ make
$ sudo make install
```

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

```
$ tar -zxvf eus_tessdata.tar.gz
$ cd  eus_tessdata
$ sudo cp eus.* /usr/local/share/tessdata
```

3.2.3 OCROPUS

Lehenik eta behin iturburu kodea jaitsi:

```
$ svn checkout http://ocropus.googlecode.com/svn/trunk/ ocropus-read-only
```

Ondoren sortu berri den karpetan sartu eta konpilatu eta instalatu daiteke:

```
$ cd  ocropus-read-only
```

OCRopusek akats bat du, parametro bezala edozein hizkuntza emanda ere beti ingelesa hartzen duela. Ondoko *patch* fitxategia aplikatuta beti euskara erabili dezan egingo dugu:

```
$ patch -p0 < ocropus_eus.patch
```

Tesseract konpilatzerakoan Hunspell erabili bada, honako pausuak jarraitu:

```
$ ../configure --without-leptonica DEFS="$DEFS
-DHAVE_HUNSPELL" LDFLAGS="$LDFLAGS -L/usr/local/lib"
```

```
$ make
$ sudo make install
```

Hunspell erabili ez bada, beste hauek:

```
$ ./configure –without-leptonica
$ make
$ sudo make install
```

3.2.4 OCROPUSGUI

Interfazea garatzeko wxWidgets eta Xml2 liburutegiak erabili dira, beraz aldez aurretik hauek instalatu beharko dira:

```
$ sudo apt-get install libwxbase2.8-dev libwxgtk2.8-dev libxml2-dev
```

OCRopusGUI eraikitzeko, jaitsi iturburu kodea *euskadinet-eko euskarazko softwarea deskargatzeko webgunetik* (http://www.euskara.euskadi.net/r59-20660/eu/contenidos/informacion/euskarazko_softwarea/eu_9567/aurkib.html) eta ondoko pausuak jarraitu:

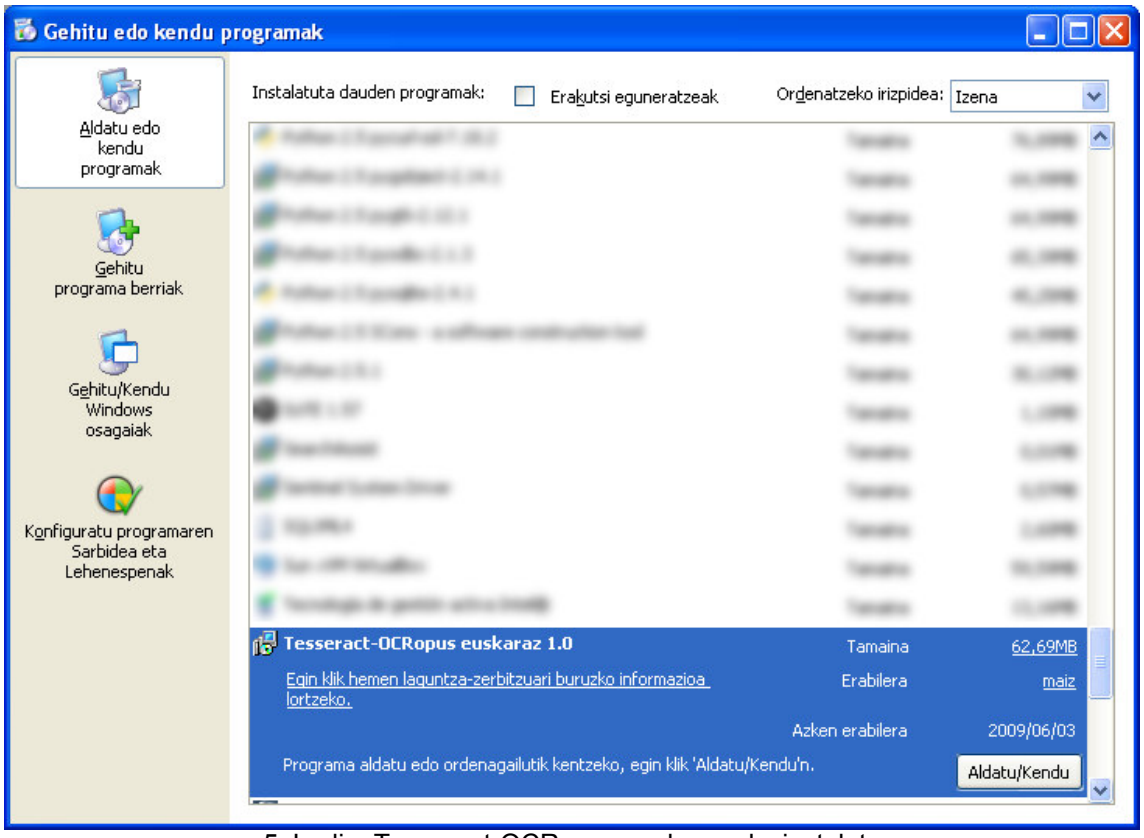
```
$ tar -zxvf ocropusgui.tar.gz
$ cd ocropusgui
$ cmake .
$ make
$ sudo make install
```

Aplikazioa martxan jartzeko ondoko agindua exekutatu:

```
$ ocropusgui
```

4 Desinstalazioa

4.1 Windows ingurunean



5. Irudia: Tesseract-OCRopus 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 Kode irekiko euskarazko OCRa aukeratu eta desinstalatzailearen pausoak jarraitzea.

4.2 Linux ingurunean

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

```
$ sudo make uninstall
```

OCRopusGUIren kasuan komando lerroz egin behar da:

```
$ sudo rm /usr/local/bin/ocropusgui
```

Iturburu kodea dagoeneko sisteman ez badago, eskuz ezabatu instalazio fitxategiak:

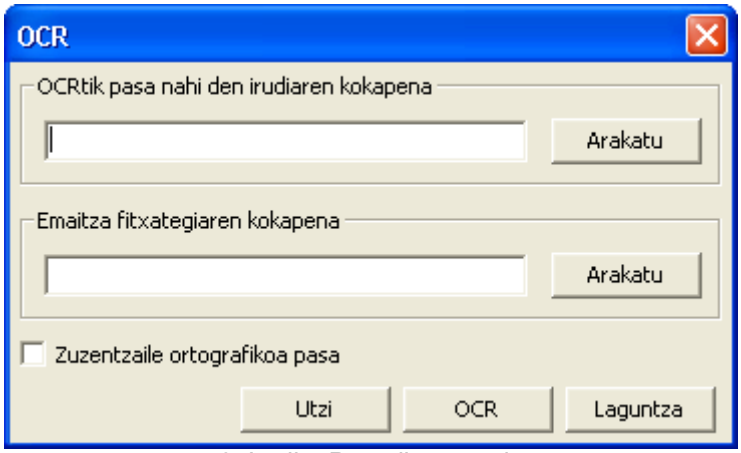
```
$ sudo rm -Rf /usr/local/include/ocropus
$ sudo rm -Rf /usr/local/include/tesseract
$ sudo rm -Rf /usr/local/include/iulib
$ sudo rm -Rf /usr/local/include/colib
$ sudo rm -Rf /usr/local/include/hunspell
$ sudo rm -Rf /usr/local/share/ocropus
$ sudo rm -Rf /usr/local/share/tessdata
$ sudo rm /usr/local/lib/libocroscript.a
$ sudo rm /usr/local/lib/libhunspell*
$ sudo rm /usr/local/lib/libocropus.a
$ sudo rm /usr/local/lib/libtesseract_*
$ sudo rm /usr/local/lib/iulib.a
$ sudo rm /usr/local/bin/ocropusgui
$ sudo rm /usr/local/bin/ocroscript
$ 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/tesseract
$ sudo rm /usr/local/bin/hunspell
```

5 Erabilera

5.1 Windows nahiz Linux ingurunean

Behin *Tesseract-OCRopus 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 zabaldu eta *ocropusgui* agindua exekutatu):



6. Irudia: Pantaila nagusia

Goiko testu kutxan irudiaren kokapen osoa idatzi beharko dugu edo ordenagailuan zehar topatu beharko dugu *Arakatu* botoia erabiliz. Sarrerako irudia .png edota .jpg motakoa izan daiteke. Beheko testu kutxan, berriz, emaitza fitxategiaren kokapena (fitxategiaren izena barne) jarriko dugu, edo Arakatu botoia erabiliz ezarriko dugu. Emaitza fitxategiaren formatua .html edo .htm motakoa izango da.

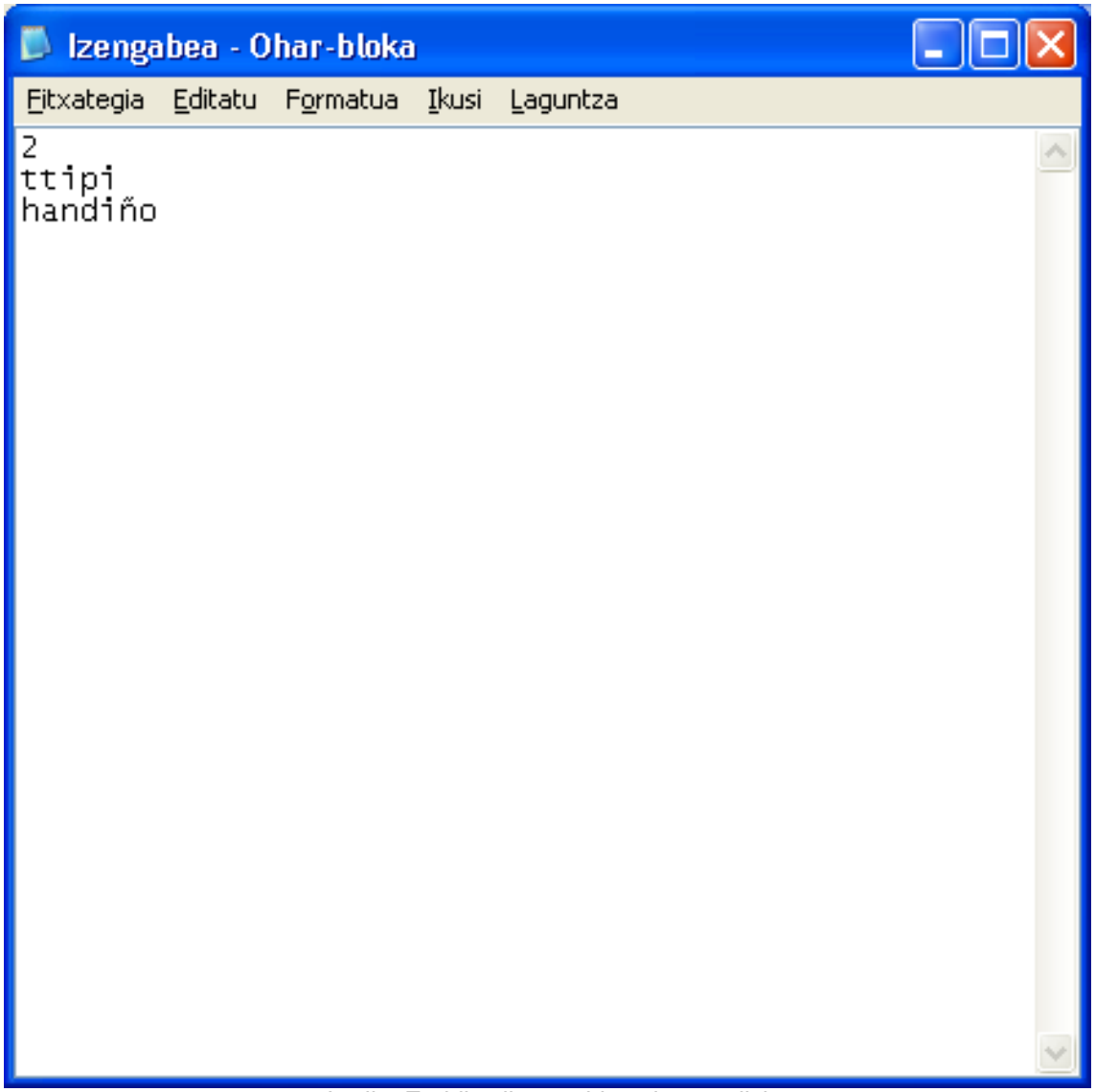
Behin sarrera eta irteera fitxategiak finkatuta ditugularik, *OCR* botoia sakatuko dugu prozesua martxan jartzeko. Prozesua amaitutakoan emaitza fitxategia irekiko da automatikoki lehenetsitako nabigatzailean.

Zuzentzaile ortografikoa pasa aukeratuz gero OCRak usuen egiten dituen erroreak zuzentzen saiatuko da aplikazioa. Adibidez, *m* letra duen hitz ezezagun bat aurkitu ezkero *m* letra *rn* letrez ordezkatzen saiatuko da hitz ezagun bat aurkitu nahian.

Laguntza botoia sakatuz dokumentu hau irekiko da.

5.2 Erabiltzailearen hiztegia

Posible da tresnaren hiztegiak hitzen bat ez ezagutzea. Kasu horretan, erabiltzaileak bere hiztegi propioa osatuz joateko aukera izango du. Horretarako, fitxategi lau bat editatzeko aukera izango du, fitxategiak ondorengo irudian azaltzen den formatua izango duelarik:



7. Irudia: Erabiltzailearen hiztegiaren edizioa

Fitxategi hau *tessdata* direktorioaren barnean aurkituko da:

- Windows ingurunean: [instalazio direktorioa]\tessdata\eus.pertsonala.dic
- Linux ingurunean: /usr/local/share/tessdata/eus.pertsonala.dic