OpenTrad: itzulpen automatiko librea, espainieratik euskarara ere

IXA taldea. Euskal Herriko Unibertsitatea

SARRERA

Dagoeneko erabilgarri daude estatu espainiarreko lau hizkuntza ofizialak (espainera (es), euskara (eu), gailegoa (gl) eta katalana (ca)) kontuan hartzen dituen *OpenTrad* proiektuaren emaitzak.

Itzulpenak transferentzia bidezko teknologia klasikoa erabiltzen duenez, hau da bi hizkuntza zehatzen arteko baliokidetzak bilatzen direnez, hizkuntza-bikote bakoitzeko sistema bat da, espainieratik euskarara (es-eu), espainieratik katalanera (es-ca), katalanetik espainierara (ca-es), espainieratik galizierara (es-gl) eta galizieratik espainierara (gl-es) bikoteetarako sistemak sortu dira. Adibideetan oinarritutako itzulpen-metodoak baztertu dira momentuz, etorkizunean teknika multzo hori modu osagarrian erantsiko bada ere. Bi teknologia sortu dira, bat *apertium* izenekoa, antz handia duten hizkuntzen artean itzultzeko; eta bestea *matxin* izenekoa, egitura desberdineko hizkuntzen artean itzultzeko. Automatetan oinarritutako teknologiari esker itzulpen-prozesua nabigazioarena bezain azkarra izan daiteke, eta, ondorioz, nabigatzen den bitartean bisitatutako orrien itzulpena lor daiteke atzerapenik gabe.

Teknologia libre eta irekia sortzea izan da helburu nagusia, beraz, estandarizazioari, modulartasunari eta hornitzaile desberdinetako moduluen elkarreragingarritasunari eman zaio lehentasuna. Proiektuaren emaitzak 2006ko martxoan publiko egin ondoren, hainbat komunikabidetan hizpide izan dira azken hilabeteetan. Emaitzak kontsultatzeko ondoko webguneak dira interesgarrienak:

- http://www.opentrad.org: sistema martxan ikusteko
- http://apertium.sourceforge.net: *apertium* teknologia eta ezagutza linguistikoa es-ca, ca-es, es-gl, gl-es itzulpenetarako eskuratzeko lizentzia librearekin
- http://matxin.sourceforge.net: *matxin* teknologia eta ezagutza linguistikoa es-eu itzulpenetarako eskuratzeko lizentzia librearekin
- http://www.blogari.net/opentrad: proiektuari buruzko bloga

Lau unibertsitateren eta hainbat enpresaren arteko elkarlanaren emaitza da proiektu hau. Unibertsitateak ondoko hauek izan dira: Euskal Herriko Unibertsitateko IXA taldea, Alacanteko Unibertsitateko Transducens taldea, Vigoko Unibertsitateko Linguistika Informatikoko Mintegia eta Kataluniako Unibertsitate Politeknikoko TALP taldea. Enpresa arduraduna Eleka Ingeniaritza Linguistikoa da, Elhuyar Fundazioaren zein Galiziako Imaxin Software enpresaren laguntzarekin. Alacanten *Prompsit* izeneko enpresa bat sortu da, proiektuaren emaitzak Herri Katalanetan zabaltzeko asmoz. Espainiako Industria, Turismo eta Merkataritza Ministerioaren laguntzaz garatu da proiektua.

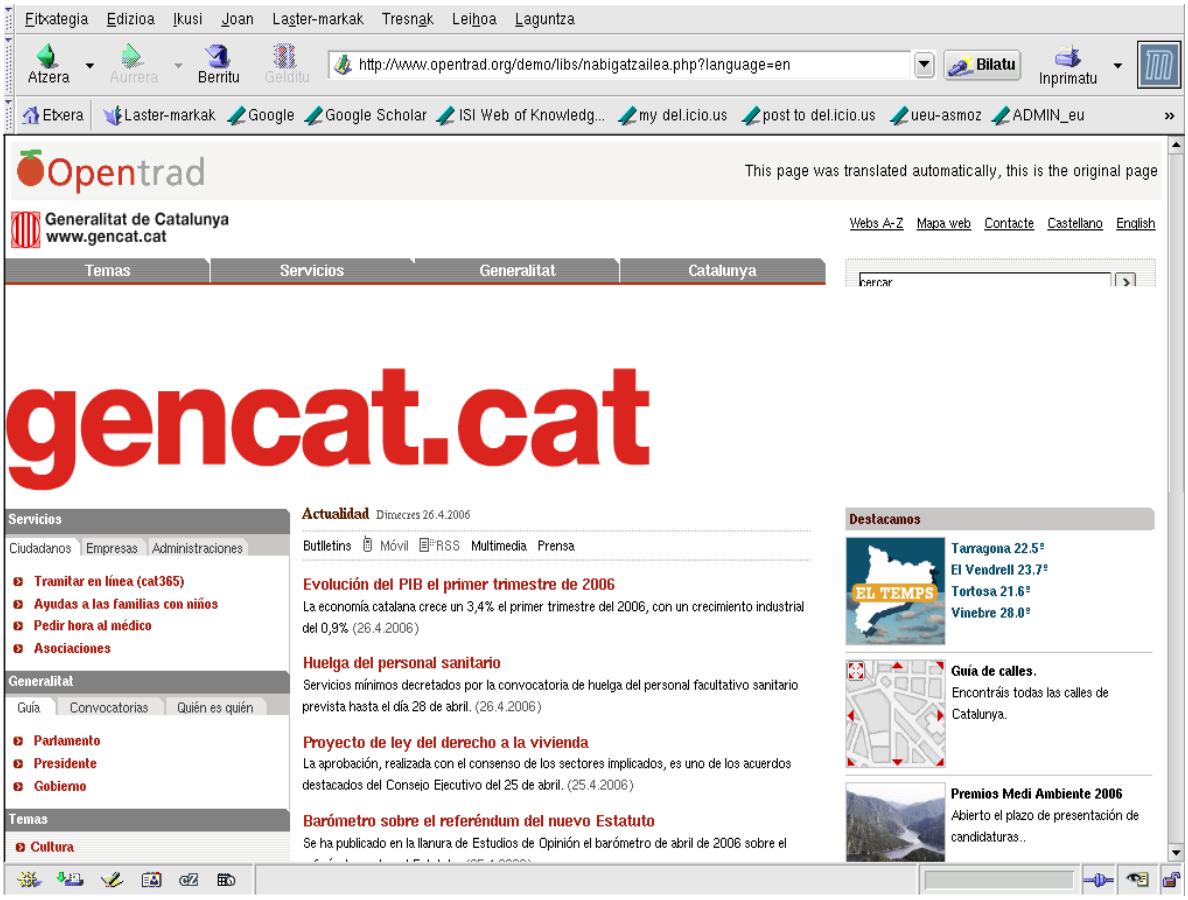
EMAITZAK

*OpenTrad*eko demoaren ondoko irudian ikus daitekeenez hiru aukera nagusi daude itzultzeko aipatutako hizkuntza bikoteen artean:

- *Testu itzulpena*: momentu horretan tekleatutako testuaren itzulpena. Probak egiteko egokia da, baina ez oso praktikoa lan profesionaletarako
- *Dokumentu itzulpena*: konputagailuan dugun dokumentu baten itzulpena. Itzulpena egiteaz gain formatuaren informazioari eusten dio dagokion tokian, beraz, lan profesionalerako aukera nagusia da. Momentuz *rtf* eta *html* formatuak dira identifikatzen eta ondo ebazten direnak, baina epe laburrean beste formatu batzuk hartu nahi dira kontuan.

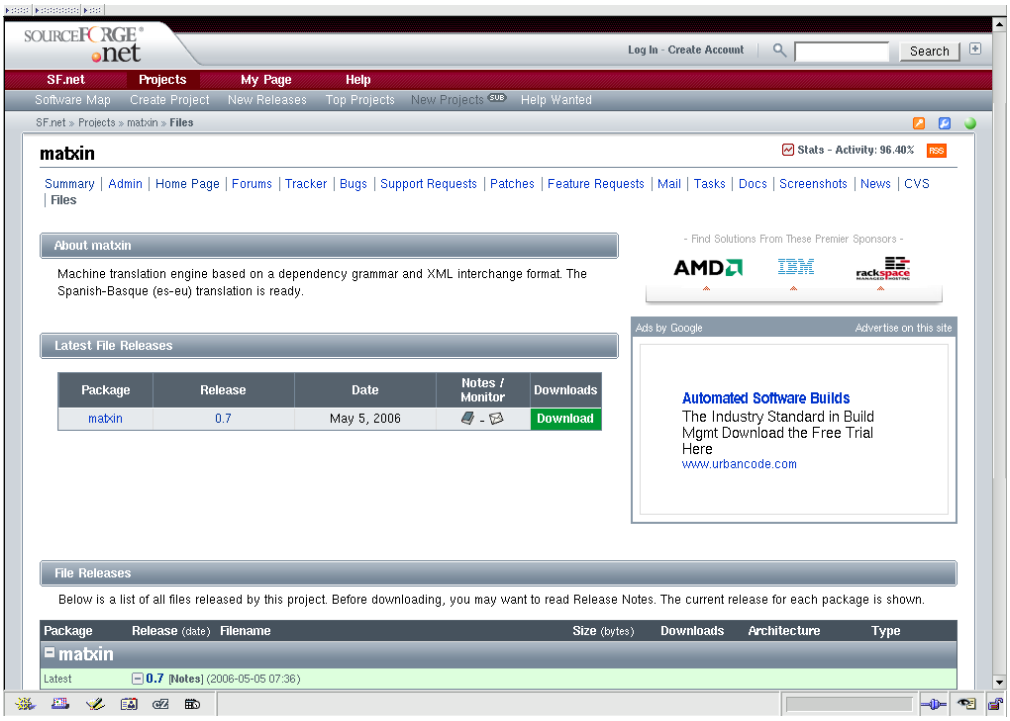
- *Nabigatu eta itzuli*: webguneak bisitatu bitartean orriak itzultzeko balio du. Zure hizkuntzan ez dauden web-orriak ikusteko oso aukera interesgarria itzulpenaren kalitatea ona baldin bada

Ondoren, azken aukera hori erabiliz, katalan hutsean dagoen web-orri baten itzulpena duzu, katalana ez dakigunentzat behintzat, aukera interesgarria:



Erabiltzeaz gain, informatikan edota hizkuntzalaritzan aditu den pertsona, erakunde edo enpresa batek teknologia eskura dezake sistema integratzeko, aldatzeko edo hobetzeko. Software libre denez, hau egitea edonoren esku dago, beti aipatutako GPL eta CC lizentzien eskakizunak betez gero (orokorrean hortik eraldatutakoa libre izan beharko dela). Aipatutako *apertium.sourceforge.net* eta *matxin.sourceforge.net* helbideetan aurkitzen dira baliabide informatikoak eta linguistikoak.

Ondoren ikus daiteke *matxin* teknologia eta es-eu baliabide linguistikoak jaisteko pantaila nagusia.



ITZULPENEN KALITATEAZ

Itzulpen automatikoan hainbat neurri erabiltzen dira lortutako kalitatea neurtzeko, eta itzultzaileentzat adierazgarriena dena honako errore-tasa hau da: hitz guztien artean zenbat aldatu behar den itzulpen zuzen bat lortzeko. Neurri hau aplikatu da sistemaren garatzaileek ezagutzen ez zuten espainieratik itzulitako esaldi multzo baten gainean eta emaitzak honako hauek izan dira:

- es-ca eta es-gl itzulpen-sistemetan errore-tasa %4 baino txikiagoa izan da, hau da, 100 hitzetatik 4 baino ez dira zuzendu behar itzulpen zuzen bat lortzeko.
- es-eu sistema, berriz, errore-tasa %32,90 izan da, hau da, hiru hitzetatik bat zuzendu behar da. Kontuan hartuta euskarazko hitz kopurua nabarmen txikiagoa dela, euskararen izaera flexibo eta eranskariarengatik, neurria normalizatu behar izan da konparagarria izan dadin eta errore-tasa konparagarria %24,80 da.

Itzulpen automatikoan errore-tasa %10 baino txikiagoa izan behar du sistema produktiboan integratu ahal izateko, beraz, es-eu sistema prototipo bat da (halaxe definituta zegoen proiektuan), oraindik garapena behar duena benetan eraginkorra izan dadin. Edozein kasutan egitura desberdineko hizkuntzen artean automatikoki itzultzean lortzen diren emaitzak neurri horretatik kanpo geratzen dira eta bi aukera daude:

- oso sistema konplexuak egitea diru handiak inbertituz
- itzulpen-sistema orokorrak alde batera utzi eta testu-mota berezi batzuetan espezializatutako sistemak sortzea

Edozein kasutan, erroreen iturburua aztertu dugu zein modulutan sakondu behar dugun detektatzeko, eta hauexek dira lortutako ondorio nagusiak:

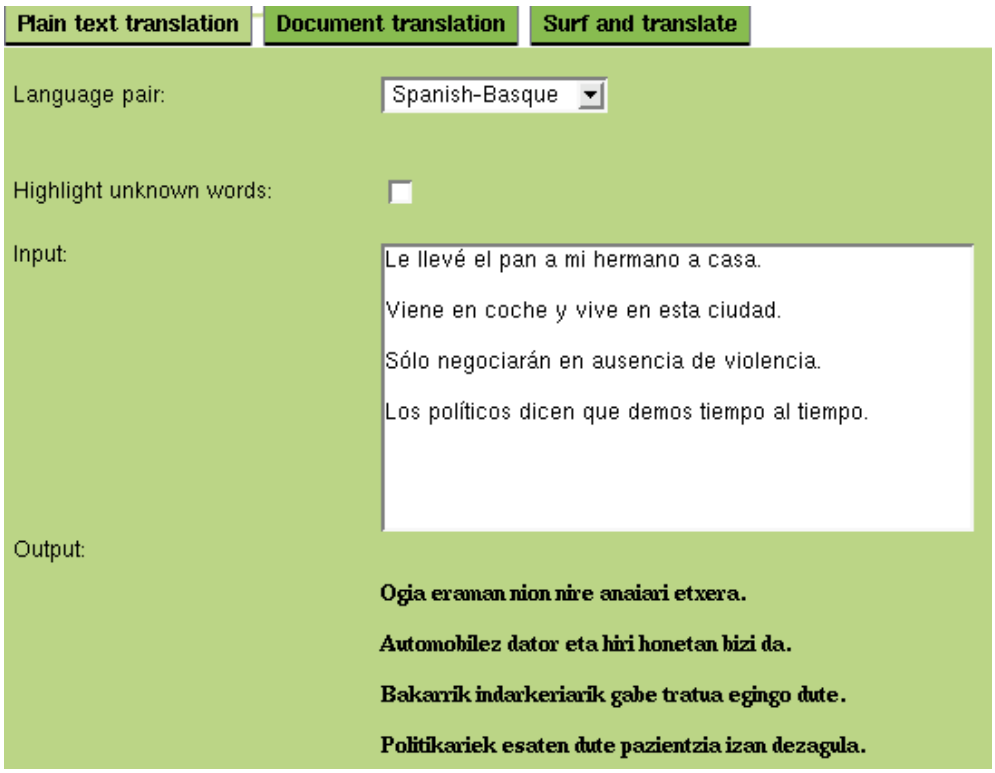
- es-ca eta es-gl itzulpenenetan arazo nagusia hiztegian ez dauden hitzak dira, izen arruntak eta izen bereziak batez ere
- es-eu itzulpenean berriz, erroreen iturri nagusia espainieraren analisi sintaktiko sakonean

egindako akatsak dira eta adiera desegokiaren hautapena zein deklinabide desegokiaren hautapena bigarren eta hirugarren arazo nagusiak dira

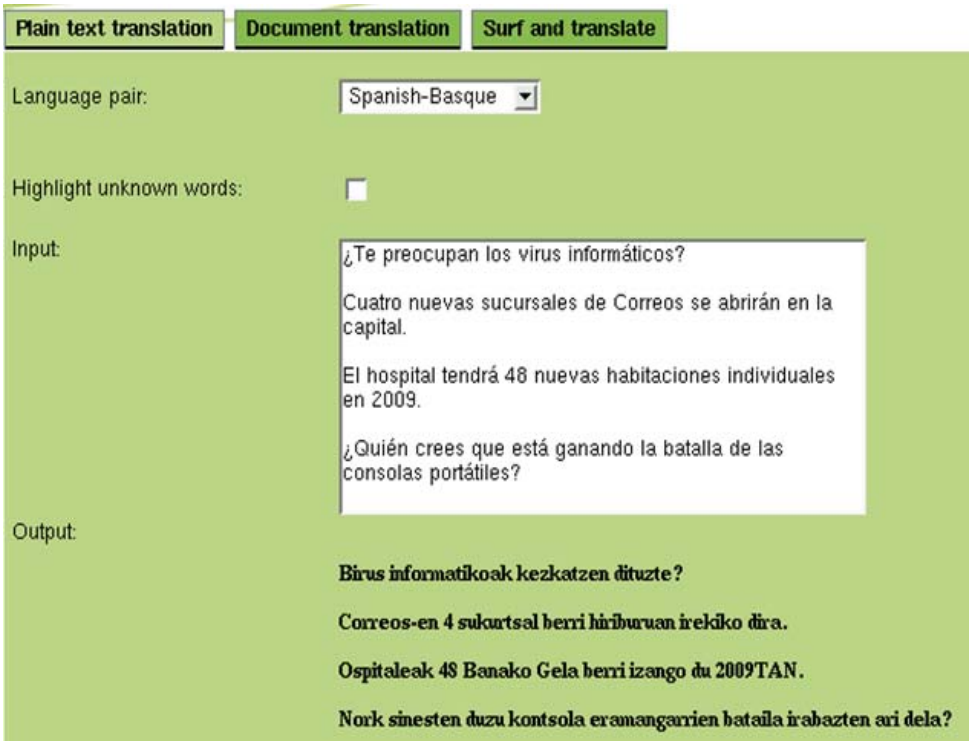
Gure asmoa espainieratik zein ingelesetik euskarara itzultzen duten sistema eraginkorrak lortzea da, baina hori lortu ahal izateko dugun teknologia itzulitako testuen ustiapenarekin konbinatu beharko da, eta hori bideragarria izan dadin aplikazio-domeinua murriztu beharko da. Gure ustez, momentu honetan ezinezkoa da itzultzaile automatiko orokor eraginkor bat eraikitzea, baina bai domeinu jakin baterako, domeinu horretan itzulitako testu asko (itzulpen-memoriak deitzen zaie) eskuragarri baldin badago.

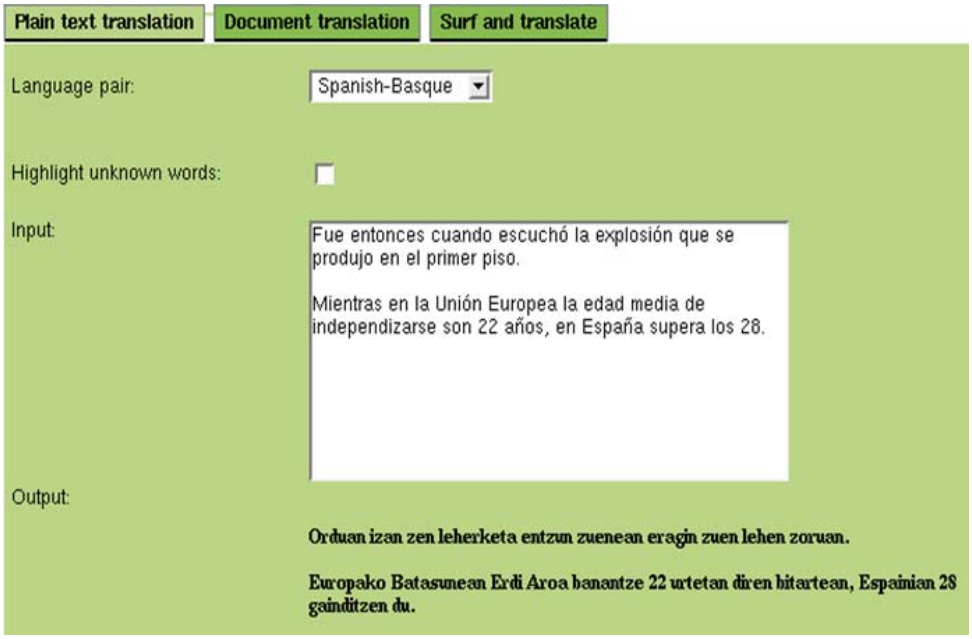
ESPAINIERA-EUSKARA ITZULPENAREN ADIBIDEAK

Proiektatu zen sistemaren helburua xumea zen, espainierazko esaldi sinpleak euskarara itzultzen duen prototipo bat garatzea. Lortutako emaitzak nahiko onak dira esaldi sinpleentzat ondoko adibidean ikus daitekeen moduan.



Dena den, esaldi horiek probatu ziren tresnaren garapenean zehar, beraz, logikoa da hauetan ondo asmatzea. Garatzaileek ezagutzen ez zituzten esaldiak dira ebaluatzeko erabili direnak eta ondoko bi iruditan adibide batzuk ikus daitezke. Aurreneko adibidean esaldien itzulpena ulergarria da, nahiz eta akats batzuk egon. Bigarrenean, berriz, esaldia desitxuratu egiten da eta jatorrizkoa begiratu gabe ez dago ulertzerik.





Azpimarra daiteke *edad media* terminoaren itzulpena: *Erdi Aroa* hautatu da testuinguru honetan itzulpen desegokia izan arren. Letra xeheek (termino moduko itzulpena aukeratzea letra larriz dagoenean) lagun dezakete akats hau konpontzen, baina antzeko kasuak gertatzen dira beste termino zein hitz batzuekin.

AKATS NABARMENENAK

Itzulpena egiteko espainieraz dagoen testuaren analisi sintaktiko osoan oinarritzen da. Analisi sintaktiko oso honek kategoria, morfologia, dependentzia sintaktikoak, etab. ditu eta itzulpena analisi horretan oinarritzen da. Beraz, analisi hori gaizki egina baldin badago, itzulpena txarra izango da.

Adibidez: *Itzul* zerrendan ezaguna egin den adibide honetan "*Miren lleva sus manzanas en un*

cesto" hau dugu itzulpena: "Haren sagarrak zaramazte saski batean begira bezate".

Espainierazko analizatzaileak "*Miren*" aginterako adizki gisa analizatzen du, 3. pertsona plurala. Hortik sortu du "*begira bezate*".

Bestetik, aditz nagusia "*Miren*" dela agertzen da, eta horregatik beste elementu guztiak bere azpian zintzilikatzen ditu zuhaitz sintaktikoan. Aditz nagusi izate horrek "*begira bezate*" hori azken lekuan agertzea dakar. Bide batez esan esaldi honetan dagoen beste errore handia "lleva" "zaramazte" gisa itzultzea dela, eta hau datu base lexikoan dugun eta konponbidean dagoen errore batek eragindako akatsa da.

Opentrad es-eu itzultzaile automatikoak duen bigarren muga handia adiera-desanbiguazioa ez egitea da. Adibidez, "*ha pasado por delante de casa*" "e*txearen aurretik atzean utzi du*" itzultzen du.

Tresnak Elhuyar-en agertzen den lehen adierako lehen hitza hartzen du eta hori erabiltzen du itzulpen gisa. "*pasar*" aditzak hainbat balizko itzulpen ditu, baina lehenengoa "atzean utzi" da. Halakoak detektatzen baditugu, itzulpenean dauden hitzen ordena aldatu egiten dugu tartean orokorrago bat dagoela iruditzen bazaigu.

Adibidez, Elhuyar Hiztegian "tráfico"ren lehen ordaina "*salerosketa*" da. *"salerosketa"* ez zaigu testuetan askotan agertzen, eta, orokorragoa den *"trafiko"*ren alde egin dugu. Oraintxe martxan dugun tresnak *"tráfico" "trafiko"* gisa itzultzen du. Badakigu zenbaitetan maileguetara jotzen ari garela, baina ez dugu uste tresnak duen akats larriena hori denik.

Hala ere, ez dugu esango inolako adiera-desanbiguaziorik egiten ez dugunik. Preposizioen kasuan badugu zenbait informazio interesgarri (hala nola izenen biziduntasuna, aditzen azpikategorizazioa,...) eta informazio hori erabiltzen saiatu gara. Adibidez, bi esaldi hauek itzultzera bidaliz gero, *"sobre"* preposizioak itzulpen ezberdinak ditu: *"el libro está sobre la mesa"* eta *"los aviones volaron sobre la muchedumbre"*. Hauek dira itzulpenak: *"liburua maihaiaren gainean dago"* eta *"Hegazkinek jendetzaren gainetik hegan egin zuten"*.

INFORMAZIO GEHIAGO

Elhuyar aldizkaria: http://www.zientzia.net/artikulua.asp?Artik_kod=11907

Norteko Ferrokarrilla irratsaioa. Hemen entzun dezakezue Iñaki Alegria eta Iñaki Arantzabali egindako elkarrizketa.

Proiektuari buruzko bloga: www.blogari.net/opentrad

Artikulu zientifiko bat (ingelesez):
http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1117456805/publikoak/es-eu-diseinua.pdf

Matxin teknologiaren dokumentazioa (espainieraz):
http://matxin.cvs.sourceforge.net/matxin/matxin/doc/