# *Matxin*, euskararako lehenengo itzultzaile automatikoa

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza,
Gorka Labaka, Mikel Lersundi, Kepa Sarasola

*Euskara, jalgi hadi plazara.*
*Lengoajetan ohi hintzan estimatze gutitan;*
*orai aldiz hik behar duk ohorea orotan.*
*Euskara, jalgi hadi mundura.*

Bernat Etxepare, 1545

## *Sarrera*

Euskarak bizirautea nahi badugu euskarak plazara jalgi behar du. Eta horrek, gaur egun, informazio-gizartearen plazara jalgi behar duela esan nahi du. Informazioaren aro honetan euskarak, beste hizkuntzen parean, nazioarteko komunikaziorako eta ulermenerako tresnak izan behar ditu.

Hizkuntzen arteko harreman horretan, gaur egungo mundu globalizatuaren Babelgo dorrean, itzulpen automatikoak (IA) garrantzi handia hartzen du eta, zenbaitetan, nahitaezko bitartekaria bihurtzen da. Euskarari dagokionez, itzulpen automatikoa beharrezkoa izango da, alde batetik, euskaldunek erdal hizkuntzetako idatziak ulertzeko, eta bestetik, euskarazko produkzioa erdaldunengana iritsi ahal izateko. Gainera, Euskal Herrian bizi ditugun errealitate elebidunetan, itzulpenen eskaera handiari aurre egiteko, itzultzaileentzako laguntza-tresnak oinarrizko lanabes izango dira, ekoizpena areagotzeko eta kostuak gutxitzeko.

Artikulu honetan *Matxin* aurkezten dugu, euskararekin lan egiten duen eta publikoki erabilgarria den lehenengo itzulpen automatikoko sistema[1].

1. Itzulpen automatikoaren eta *Matxin* sistemaren eraikuntzaren inguruan sakondu nahi duenak zehaztasun guztiak aurkituko ditu Aingeru Mayor-ek 2007. urtean aurkeztutako "Matxin. Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz" doktoradutza-tesian: http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1196444990/publikoak/Tesi txostena.pdf

Ingeniaritza linguistikoko ikerketa-lan hau IXA[2] taldearen barruan garatua izan da, talde honek euskararen prozesamendurako diseinatutako estrategiaren barruan.

Azken hamarkadako joera, IAren arloan, corpusetan oinarritutako estrategiak erabiltzea izan da, erregeletan oinarritutako lanak gutxietsiz. Baina, euskarara automatikoki itzultzeko hurbilpen estatistikoen erabilerak zailtasun handiekin topo egiten du. Batetik, teknika estatistikoek corpus erraldoiak behar dituzte emaitza onargarriak lortzeko eta euskararako gaur egun eskuragarri dauden corpusak (eta etorkizunean egongo direnak ere) mugatuak dira; bestetik, euskara hizkuntza eranskaria da eta morfologia aberatsa duten hizkuntzetara itzultzean sistema estatistikoak erregeletan oinarritutako sistema komertzialen atzean gelditzen dira.

Etorkizuna hibridazioaren bidetik etorriko da seguruen, eta gaur egun dauzkagun datuekin, aurreikus dezakegu euskarara itzultzeko sistema hibrido horietan erregeletan oinarritutako teknikek pisu handia izango dutela. Beraz, gure helburua erregeletan oinarritutako estrategiaren ahalmena aztertzea izan da, hori bai, gaur egungo erronkei aurre eginez: berrerabilgarritasuna, estandarizazioa eta kode irekia.

Gure lanaren fruitua *Matxin* da, transferentzia sintaktiko sakona egiten duen erregeletan oinarritutako IAko sistema. Hainbat tresna eta baliabide linguistiko berrerabili ditugunez, eta etorkizunean beste hizkuntzetara hedatzea eta beste moduluak integratzea aurreikusten dugunez, formatuen estandarizazioa beharrezkoa izan da elkarreragingarritasuna bermatzeko.

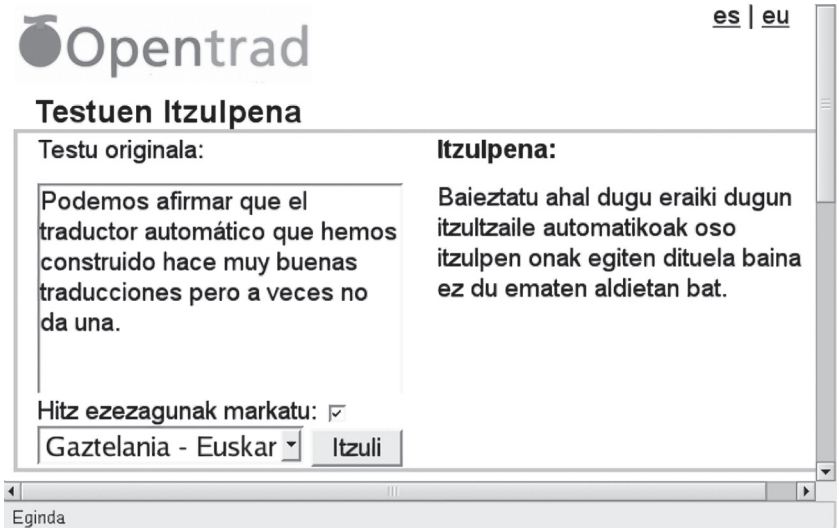
Sistemaren arkitektura abiapuntu- eta xede-hizkuntzetatik independentea izateko diseinatua izan da. Espainieratik euskarara itzultzen duen *Matxin 1.0* prototipoa erabilgarri dago Interneten[3] (ikus 1. irudia) eta kode irekiko software libre bezala banatzen da[4]. Une honetan sistema hedatzen ari gara ingelesetik euskarara ere itzul dezan.

Ondokoa da artikulu honen egitura. 2. atalean *Matxin* itzulpen-sistemaren ezaugarriak eta arkitektura orokorra deskribatzen ditugu. Hurrengo hiru ataletan analisia, transferentzia eta sorkuntza faseetako moduluak aurkezten ditugu. 6. atalean gure sistemaren ebaluazioaren emaitzak argitaratzen ditugu, eta azkenik ondorioak eta etorkizunerako lanak laburbiltzen ditugu.

## Matxin *itzulpen-sistemaren teknologia*

IXA taldeak oinarri linguistikoak ezarri eta konplexutasun ertaineko tresnak garatu ondoren, tresna aurreratuen inguruko ikerketari ekin zion, eta 1998. urtean erabaki zuen itzulpen automatikoaren (IA) erronkari aurre egitea. Lehenengo prototipoak izen- eta preposizio-sintagmak ingelesetik euskarara itzultzen zituen. Ondoren espainieratik euskarara itzultzen zuen prototipoa inplementatu genuen, esaldi mailara jauzia emanez.

2. http://ixa.si.ehu.es
3. http://www.opentrad.org
4. http://matxin.sourceforge.net




Irudia 1: *Matxin 1.0* prototipoa Interneten

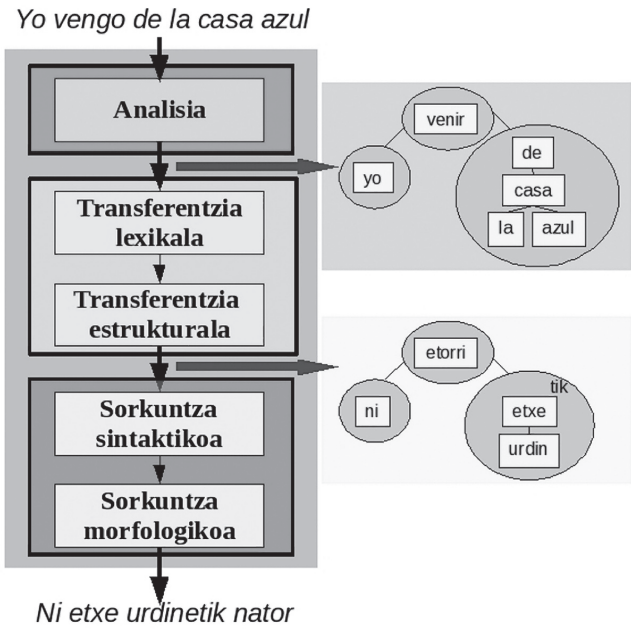
2005.ean OpenTrad[5] proiektua martxan jarri zen. Proiektu horren helburua estatu espainiarreko hizkuntza nagusietarako abiadura handiko eta kode irekiko itzulpen automatikoko sistemak sortzea zen. Aurretik garatutako bi gailu hobetu eta integratu ziren: *Apertium*[6], gertuko hizkuntza bikoteentzat (espainiera, galegoa eta katalana) transferentzia sintaktiko partziala egiten duen kode irekiko sistema arrakastatsua (Corbí-Bellot *et al.*, 2005), eta *Matxin*, elkarrengandik urrunago dauden hizkuntza bikoteentzat (espainiera-euskara) transferentzia sintaktiko sakona burutzen duena.

2007an *Matxin* sistemaren erroreak aztertu eta hobekuntza ugari egin ondoren, *Matxin 1.0* bertsio zenbakiarekin banatu egin genuen. Gaur egun hobekuntzak EurOpenTrad proiektuaren barruan egiten ari dira.

## 1. Ezaugarriak

*Matxin* sistemaren arkitektura (2. irudia) transferentzian oinarritutako eredu klasikoan oinarritzen da. Itzulpen-prozesua hiru fasetan banatzen da analisia (ikus 3. atala), transferentzia (4. atala) eta sorkuntza (5. atala). Gainera dokumentuen formatua mantentzeko eta postediziorako moduluak gehitu dira.

5. http://www.opentrad.org
6. http://www.apertium.org




Irudia 2: Arkitektura orokorraren eskema.

Fase bakoitza hainbat modulutan banatzen da eta itzulpen-prozesuko ataza linguistikoek gidatu dute moduluen arteko banaketaren diseinua. Datu linguistikoak eta erregelak algoritmoetatik kanpo daude. Modulu elebakarrak modulu elebidunetatik ahalik eta independenteenak dira eta sistema erdal hizkuntzetatik euskarara itzultzeko diseinatuta dago, abiapuntu-hizkuntzatik independentea izanik. Moduluen arteko komunikaziorako interfaze egokia finkatzeak berezko garrantzia duenez, datuen egitura arreta handiz diseinatu dugu.

Berrerabilgarritasuna izan da sistemaren eraikuntzaren gakoa. Aurretik eraikitako moduluak (espainierazko analizatzailea, euskararako sortzaile morfologikoa, desformatatzailea eta birformatatzailea…) eta sortutako baliabide linguistikoak (hiztegiak eta corpusak) berrerabili ditugu. Gainera guk sortutako moduluak eta datu linguistikoak berrerabilgarri izan daitezen eraiki ditugu: espainierazko mendekotasun-analizatzailea, preposizioen hiztegia, aditz-kateen transferentzia, etab.

Berrerabiltzeak dakarren ondorioetako bat baliabideen eta moduluen hetereogeneotasuna denez, modulu berrerabilien arteko elkarreragingarritasuna ziurtatzea nahitaezkoa da. Honetarako oinarrizkoa izan da moduluen arteko datuen fluxuak, baliabide linguistikoek eta datu-egitura manipulatzeko erregelen formalismoak formatu estandar bati jarraitzea, XML[7] aukeratu

7. *Extensible Markup Language*. http://www.w3.org/XML

dugula. Hiztegiak *Apertium* (Forcada *et al.*, 2006) proiektuaren espezifikazioari jarraitzen dioten XMLn oinarritutako formatuan kodetuta daude; proiektu horretako konpiladore batekin hiztegi horiek fitxategi bitarrak bihurtzen ditugu, oso azkar prozesatuko diren egoera finituetako transduktoreetan oinarritutako adierazpide bat sortuz. Datu-egiturak manipulatzeko erregelen formalismoa XPath[8] lengoaian oinarritzen da eta aditz-kateen transferentziarako gramatika XFST[9] tresna zabalduaren sintaxian.

## 2. Esaldiaren itzulpena prozesatzeko datu-egitura

Gure sistemak prozesatuko duen datu-egitura esaldien egitura sintaktikoan oinarritzen da. Egitura sintaktikoa bi modutan adieraz daiteke (Civit, 2003): osagaiekin edo mendekotasunekin. Gure sistemarako, transferentzia sakonerako beharrezkoak diren bi formalismoen ezaugarriak biltzen dituen egitura sintaktiko hibrido bat proposatzen dugu: osagaiak etiketatzen dira eta osagai bakoitzeko hitzen arteko eta osagaien arteko mendekotasun-erlazioak ere adierazten dira.

Datu-egiturak hiru objektu-mota erabiltzen ditu: esaldia, chunka eta nodoa. Esaldia itzulpenerako unitatea da; sarrerako testua esaldietan banatu ondoren, esaldiak banan-banan itzultzen dira, beste esaldiak kontutan hartu gabe. Chunka osagai bat adierazten duen sasi-sintagma ez-errekurtsiboa da; gure sisteman chunkek garrantzi handia dute prozesamendua errazteko, modulu bakoitzak maila bakar batean lan egiten duelako, chunk barruan edo chunken artean. Nodoak hitz bat edo hitz anitzeko unitate bat adierazten du. Mendekotasun-zuhaitzean esaldi bakoitzaren menpe chunk erroa dago; chunk bakoitzaren menpe chunk horretako nodo erroa eta chunk horren menpeko chunkak daude; eta nodo bakoitzaren menpe nodo horren azpiko nodoak.

Diseinatu dugun datu-egitura hau transferentzia- eta sorkuntza-faseetako moduluek prozesatuko dute, eta moduluen arteko komunikaziorako erabiliko da. Eragingarritasuna bermatzeko XMLn oinarritutako datu-egitura arin bat diseinatu dugu, esaldi bakoitzaren itzulpenerako beharrezkoa den informazio guztia edukitzeko ahalmenarekin. DTDan (ikus 3. irudia) itzulpen-prozesuko hiru elementu nagusiak (esaldiak, chunkak eta nodoak), beren atributuak eta mendekotasun-erlazioak deskribatzen dira. Elementuen atributuek informazio linguistikoa edo formatukoa adierazten dute.

Datu-egituraren formatuko atributuen erabilera hobeto ulertzeko 2.3 atalean adibide baten itzulpen-prozesua deskribatuko dugu.

8. *XML Path Language*. http://www.w3.org/TR/xpath

9. *Xerox Finite-State Tool*. http://www.cis.upenn.edu/ cis639/docs/xfst.html

```
<!ELEMENT SENTENCE (CHUNK+)>
 <!ATTLIST SENTENCE            <!--Esaldia-->
    ord        CDATA #IMPLIED  <!--Esaldiaren ordena testuan-->
    ref        CDATA #IMPLIED  <!--XHko esaldiari dagokion AHkoa-->
    alloc      CDATA #IMPLIED  <!--1.hizkiaren posizioa testuan-->
 >
 <!ELEMENT CHUNK (NODE, CHUNK*)>
 <!ATTLIST CHUNK               <!--Chunka-->
    ord        CDATA #IMPLIED  <!--Chunkaren ordena esaldian-->
    ref        CDATA #IMPLIED  <!--XHko chunkari dagokion AHkoa-->
    alloc      CDATA #IMPLIED  <!--1.hizkiaren posizioa testuan-->
    type       CDATA #IMPLIED  <!--Chunk-mota-->
    si         CDATA #IMPLIED  <!--Funtzio sintaktikoa-->
    focus      CDATA #IMPLIED  <!--Fokua-->
    prep       CDATA #IMPLIED  <!--Preposizioa-->
    trans      CDATA #IMPLIED  <!--Iragankortasuna-->
    subper     CDATA #IMPLIED  <!--Subjektuaren pertsona-->
    <!...>
 >
 <!ELEMENT NODE (NODE*)>
 <!ATTLIST NODE                <!--Nodoa-->
    ord        CDATA #IMPLIED  <!--Nodoaren ordena chunkean-->
    ref        CDATA #IMPLIED  <!--XHko chunkari dagokion AHkoa-->
    alloc      CDATA #IMPLIED  <!--1.hizkiaren posizioa testuan-->
    form       CDATA #IMPLIED  <!--Forma-->
    lem        CDATA #IMPLIED  <!--Lema-->
    mi         CDATA #IMPLIED  <!--Informazio morfologikoa-->
    pos        CDATA #IMPLIED  <!--Part-of-speech: kat. eta azpikat.-->
    suf        CDATA #IMPLIED  <!--Atzizkiaren informazioa-->
    det        CDATA #IMPLIED  <!--Mugatasuna-->
    num        CDATA #IMPLIED  <!--Numeroa-->
    per        CDATA #IMPLIED  <!--Pertsona-->
    loc        CDATA #IMPLIED  <!--Kokapen-informazioa-->
    lmi        CDATA #IMPLIED  <!--Lemaren segmentazio morfologikoa-->
    sem        CDATA #IMPLIED  <!--Informazio semantikoa-->
    post       CDATA #IMPLIED  <!--Postposizioa-->
    spost      CDATA #IMPLIED  <!--Menpeko postposizioa-->
    vpost      CDATA #IMPLIED  <!--Aditz-postposizioa-->
    prep       CDATA #IMPLIED  <!--Preposizioa-->
 >
```

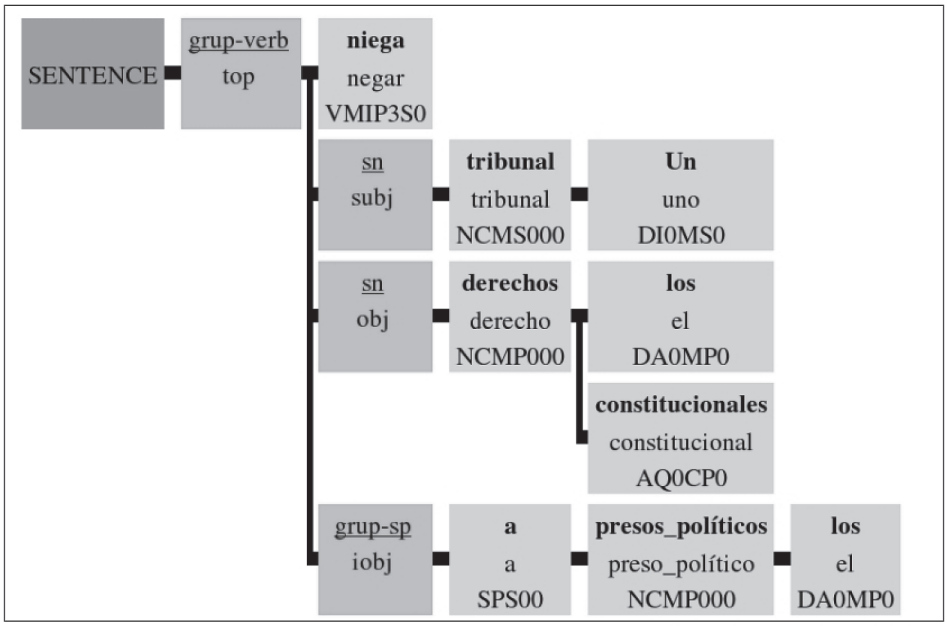
Irudia 3: Esaldien itzulpena prozesatzeko datu-egitura: DTDa

## 3. Itzulpen-prozesua: adibide bat

*Matxin* sistemaren arkitektura eta datu-egitura hobeto ulertzeko itzulpen-prozesua burutzen duten moduluen arteko datu-fluxua erakutsiko dugu ondoko adibidea erabiliz: *Un tribunal niega los derechos constitucionales a los presos políticos.*

Analisi-fasearen irteera 4. irudian ikus dezakegu. Elementu bakoitzaren ordena ord atributuan ikusten da. Jatorrizko testuan nodo bakoitzaren lehenengo hizkia okupatzen duen posizioa (*alloc*) dokumentuaren formatua berreskuratzeko erabiliko da. Bestelako atributuak itzulpen-prozesurako beharrezkoak diren sarrerako chunk eta nodoen informazio linguistikoa adierazten dute. Informazio morfologikoak (*mi*) *Parole* notazioari jarraitzen dio.

Transferentzia lexikalaren ondoren (5. irudia) honako aldaketak ikus ditzakegu: *ord* atributuan zegoena *ref* atributuan gordetzen da jatorrizko testuaren ordenaren erreferentzia erabili

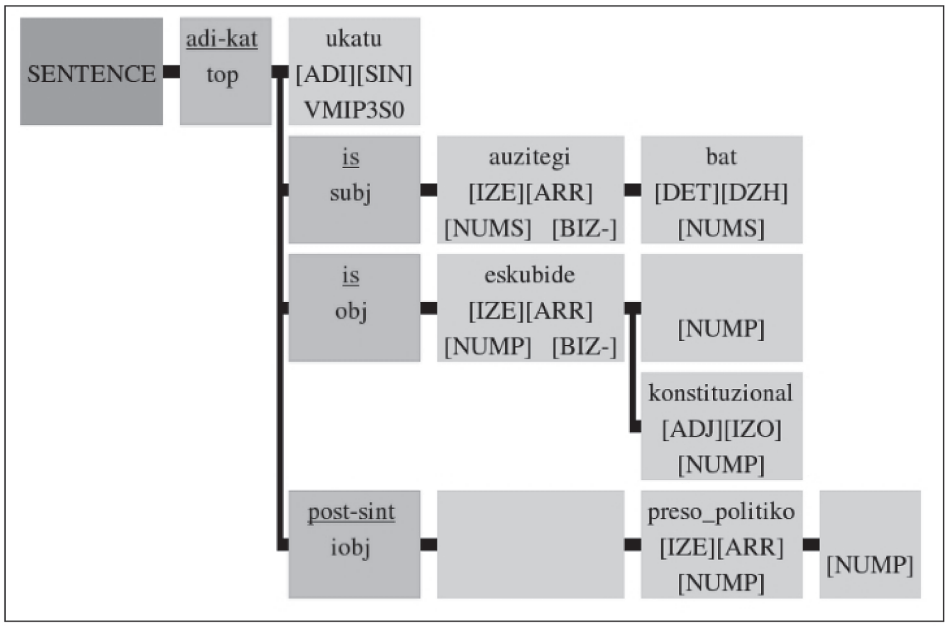



```
<SENTENCE ord='1' alloc='0'>
 <CHUNK ord='2' alloc='12' type='grup-verb' si='top'>
  <NODE ord='1' alloc='12' form='niega' lem='negar' mi='VMIP3S0'>
  <CHUNK ord='1' alloc='0' type='sn' si='subj'>
   <NODE ord='2' alloc='3' form='tribunal' lem='tribunal' mi='NCMS000'>
    <NODE ord='1' alloc='0' form='Un' lem='uno' mi='DI0MS0'/> </NODE> </CHUNK>
  <CHUNK ord='3' alloc='18' type='sn' si='obj' focus='true'>
   <NODE ord='2' alloc='22' form='derechos' lem='derecho' mi='NCMP000'>
    <NODE ord='1' alloc='18' form='los' lem='el' mi='DA0MP0'/>
    <NODE ord='3' alloc='31' form='constitucionales' lem='constitucional' mi='AQ0CP0'/></NODE>
  </CHUNK>
  <CHUNK ord='4' alloc='48' type='grup-sp' si='iobj'>
   <NODE ord='1' alloc='48' form='a' lem='a' mi='SPS00'>
    <NODE ord='3' alloc='54' form='presos_polticos' lem='preso_poltico' mi='NCMP000'>
     <NODE ord='2' alloc='50' form='los' lem='el' mi='DA0MP0'/> </NODE></NODE></CHUNK></CHUNK>
</SENTENCE>
```

Irudia 4: Analisi-fasearen irteera

ahal izateko postedizioan. *ord* eta *form* atributuak desagertzen dira, sorkuntza-fasean xede-hizkuntzako testurako ordena berria kalkulatu eta hitzen forma eman beharko delako. Lexikoi elebidunean kontsulta eginez abiapuntuko nodoetako *lem* eta *mi* atributuen transferentzia lexikalak xede-hizkuntzako *lem, pos det, num* eta *sem* atributuen balioak ematen ditu. Chunketako *type* atributua ere itzuli egiten da.

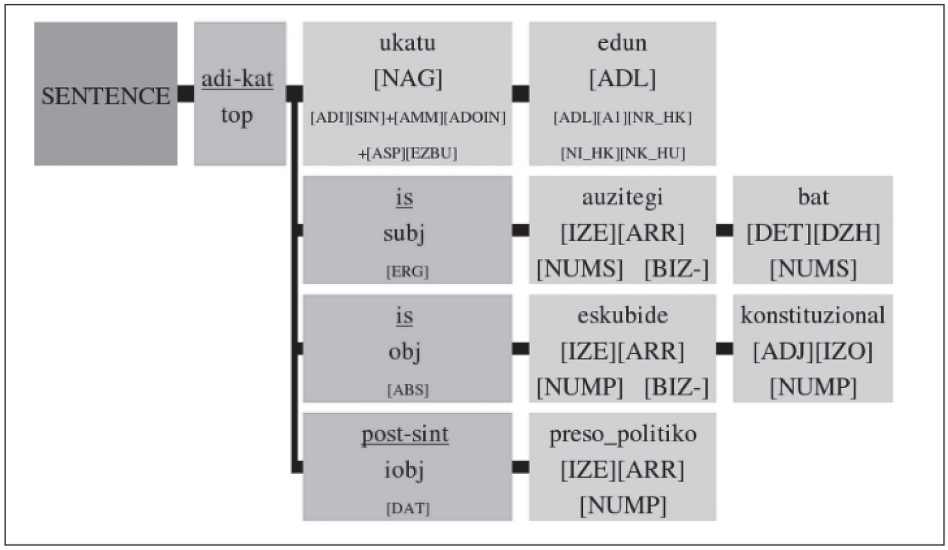
Transferentzia estrukturalean (6. irudia) balio lexikalik gabeko nodoak (preposizioak eta artikuluak) desagertzen dira, zeukaten mugatasunaren, numeroaren eta preposizioaren infor-




```
<SENTENCE ref='1' alloc='0'>
 <CHUNK ref='2' alloc='12' type='adi-kat' si='top'>
  <NODE ref='1' alloc='12' lem='ukatu' mi='VMIP3S0' pos='[ADI][SIN]'/>
  <CHUNK ref='1' alloc='0' type='is' si='subj'>
   <NODE ref='2' alloc='3' lem='auzitegi' pos='[IZE][ARR]' num='[NUMS]' sem=[BIZ-]>
    <NODE ref='1' alloc='0' lem='bat' pos='[DET][DZH]' num='[NUMS]'/> </NODE> </CHUNK>
  <CHUNK ref='3' alloc='18' type='is' si='obj' focus='true'>
   <NODE ref='2' alloc='22' lem='eskubide' pos='[IZE][ARR]' num='[NUMP]' sem='[BIZ-]'>
    <NODE ref='1' alloc='18' det='[MUGM] num='[NUMP]'/>
    <NODE ref='3' alloc='31'  lem='konstituzional' pos='[ADJ][IZO]' num='[NUMP]'/></NODE></CHUNK>
  <CHUNK ref='4' alloc='48' type='post-sint' si='iobj'>
   <NODE ref='1' alloc='48' prep='a'>
    <NODE ref='3' alloc='54'  lem='preso_politiko' pos='[IZE][ARR]' num='[NUMP]'>
     <NODE ref='2' alloc='50' det='[MUGM] num='[NUMP]'/> </NODE></NODE></CHUNK></CHUNK>
</SENTENCE>
```

Irudia 5: Transferentzia lexikalaren irteera

mazioa chunkari pasatuz. Gainera, desagertu diren nodo horien erreferentziak (*ref* eta *alloc*) chunkari ere pasatuko zaizkio (*postref* eta *postalloc* atributuetan) postedizioan eta birformatatzailean beharrezkoak izango direlako. Preposizio eta funtzio sintaktikoen transferentzia egiten da, chunketako postposizioa (*post*) lortuz. Aditz-kateen transferentzia burutzean, egitura berri bat duen chunka sortzen da, nodo bakoitzari esleitzen zaiola ordenaketarako beharrezkoa duen informazioa (*loc*), eta sorkuntza morfologikorakoa (*post*).

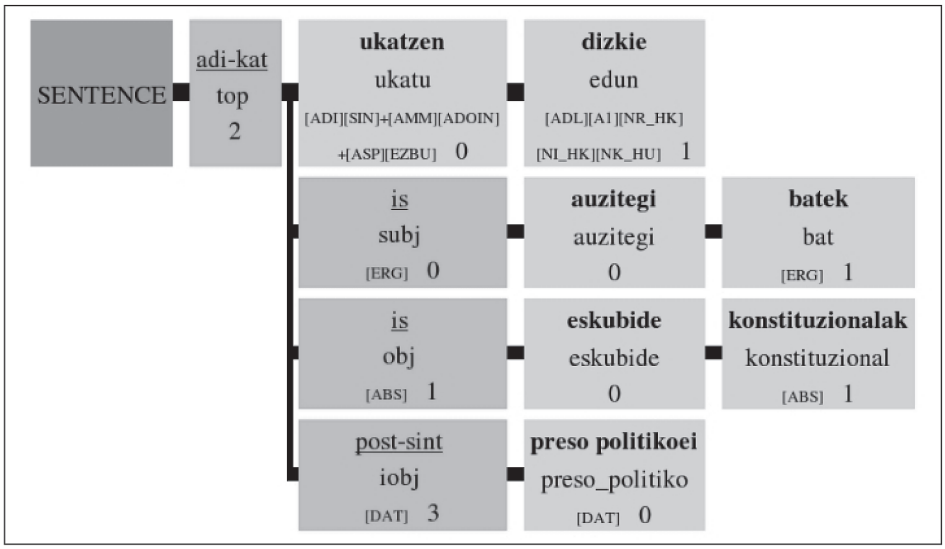



```
<SENTENCE ref='1' alloc='0'>
 <CHUNK ref='2' alloc='12' type='adi-kat' si='top'>
  <NODE ref='1' alloc='12' lem='ukatu' loc='[NAG]'
        post='[ADI][SIN]+[AMM][ADOIN]+[ASP][EZBU]'>
   <NODE ref='1' alloc='12' lem='edun' loc='[ADL]'  post='[ADL][A1][NR_HK][NI_HK][NK_HU]'/></NODE>
  <CHUNK ref='1' alloc='0' type='is' si='subj' num='[NUMS]' post='[ERG]'>
   <NODE ref='2' alloc='3' lem='auzitegi' pos='[IZE][ARR]' num='[NUMS]'>
    <NODE ref='1' alloc='0' lem='bat' pos='[DET][DZH]' num='[NUMS]'/> </NODE></CHUNK>
  <CHUNK ref='3' type='is' alloc='18'  postref='1' postalloc='18' si='obj' focus='true'
         det='[MUGM] num='[NUMP]' post='[ABS]'>
   <NODE ref='2' alloc='22' lem='eskubide' pos='[IZE][ARR]' num='[NUMP]'>
    <NODE ref='3' alloc='31' lem='konstituzional' pos='[ADJ][IZO]' num='[NUMP]'/> </NODE></CHUNK>
  <CHUNK ref='4' alloc='48' postref='1,2' postalloc='48,50' type='post-sint' si='iobj'
         det='[MUGM] num='[NUMP]' post='[DAT]'>
   <NODE ref='3' alloc='54' lem='preso_politiko' pos='[IZE][ARR]' num='[NUMP]'/></CHUNK> </CHUNK>
</SENTENCE>
```

Irudia 6: Transferentzia estrukturalaren irteera

Sorkuntzaren ondoren (7. irudia) kalkulatutako ordena berria *ord* atributuan agertzen da. Chunkeko azken hitzari sorkuntza morfologikoa burutzeko postposizio-informazioa (post) pasatzen zaio eta informazio horrekin lotura zuten desagertutako nodoen erreferentziak (*postref* eta *postalloc*). Nodo bakoitzaren *form* atributuan xede-hizkuntzako hitz-forma gordetzen da prozesadore morfologikoa erabiliz lortu dena nodoak postposizio-informazioa bazuen. Bestela, forma zuzenean lemaren informaziotik (*lem*) jasotzen da.

Adibide honetarako sistemak ematen duen irteera hauxe da: *Auzitegi batek eskubide konstituzionalak ukatzen dizkie preso politikoei.*




```
<SENTENCE ord='1' ref='1' alloc='0'>
 <CHUNK ord='3' ref='2' type='adi-kat' alloc='12' si='top'>
  <NODE ord='1' ref ='1' alloc ='12' form='ukatzen' lem='ukatu' pos='[NAG]'
        post='[ADI][SIN]+[AMM][ADOIN]+[ASP][EZBU]'>
   <NODE ord='2' ref ='1' alloc ='12' form='dizkie' lem='edun' pos='[ADL]'
         post='[ADL][A1][NR_HK][NI_HK][NK_HU]'/> </NODE>
  <CHUNK ord='1' ref='1' alloc='0' type='is' si='subj' post='[ERG]'>
   <NODE ord='1' ref ='2' alloc ='3' form='auzitegi' lem='auzitegi' pos='[IZE][ARR]'
         num='[NUMS]'>
    <NODE ord='2' ref ='1' alloc ='0' form='batek' lem='bat' pos='[DET][DZH]'
          num='[NUMS]' post='[ERG]'/> </NODE> </CHUNK>
  <CHUNK ord='2' ref='3' alloc='18' type='is' si='obj' focus='true' det='[MUGM]
         num='[NUMP]' post='[ABS]' postref='4' postalloc='18'>
   <NODE ord='1' ref ='2' alloc ='22' form='eskubide' lem='eskubide' pos='[IZE][ARR]'
         num='[NUMP]'>
    <NODE ord='2'  ref ='1,3' alloc ='31,18' form='konstituzionalak' lem='konstituzional'
          pos='[ADJ][IZO]' num='[NUMP]' post='[ABS]'></NODE></NODE></CHUNK>
  <CHUNK ord='4' ref='4' alloc='48' type='post-sint' si='iobj' det='[MUGM] num='[NUMP]'
         post='[DAT]' postref='1,2' postalloc='48,50'>
   <NODE ord='1 ref ='1,2,3' alloc ='54,48,50' form='preso politikoei'
         lem='preso_politiko' pos='[IZE][ARR]' num='[NUMP]' post='[DAT]'></NODE></CHUNK></CHUNK>
</SENTENCE>
```

Irudia 7: Sorkuntza-fasearen irteera

### *Analisia*

Analisiaren emaitza abiapuntu-testuaren errepresentazio abstraktua izango da. Gure kasuan ez du zentzurik behar dugun analizatzailea guk geuk eraikitzeak, dagoeneko espainiera analizatzeko tresna sendoak egon badaudelako. Hala ere, ikerketa hau burutzen ari zen garaian, analisi partziala soilik ematen zuten analizatzaileak zeuden eskuragarri eta guk, transferentzia sakona burutu ahal izateko, hitz bakoitzaren informazio morfologikoa lortzeaz eta chunkak identifikatzeaz gain, hitzen arteko eta chunken arteko mendekotasun-erlazioak eta funtzio sintaktikoak ematen dituen analisi osoa behar dugu.

Gure aukera UPC[10] unibertsitatean espainierarako garatutako *FreeLing* (Atserias *et al.*, 2006) analizatzaile partziala berrerabiltzea izan da, batetik, software librea delako[11] eta, bestetik, aukera ematen digulako, *Opentrad* bezalako proiektuen barnean, UPCko taldearekin elkarlanean aritzeko, analizatzailea gure beharretara egokituz. Analizatzaile partzialak esaldiko hitz bakoitzerako bere forma, lema eta informazio morfologiko desanbiguatua emateaz gain, hitzak chunketan multzokatzen ditu, chunka bere motarekin etiketatuz.

Mendekotasun-erlazioak lortzeko, analisi partzialetik abiatuta, chunken arteko eta chunk barruko nodoen arteko loturak, eta funtzio sintaktikoak ebazten dituen modulu bat diseinatu eta garatu dugu. Guk eraikitako moduluak (batez ere chunken arteko mendekotasunak ebazten dituena) bere fruituak eman ditu: hasiera batean gure sistema martxan jarri ahal izateko nahitaezko elementua izan zen eta, ondoren, *Freeling*eko garatzaileak eurak ere gure moduluan oinarritu ziren *Freeling*erako *Txala* izeneko mendekotasun-analizatzailea eraikitzeko (Atserias *et al.*, 2005).

### *Transferentzia*

Transferentziaren helburua abiapuntu-hizkuntzako testuaren adierazpide abstraktua xede-hizkuntzako adierazpidea bihurtzea da. Transferentzia bi mailatan burutzen da: lexikala eta estrukturala.

#### 1. Transferentzia lexikala

Transferentzia lexikalaren muina lexikoian bilatzea da, abiapuntu-hizkuntzako nodoen lema eta informazio morfologikoa erabiliz. Sarrera horri dagozkion xede-hizkuntzako ordain guztien lema, part-of-speech, kokapenari buruzko informazioa, pertsona eta numeroa, informazio semantikoa, ordainaren osaketa morfologikoa eta beste informazio batzuk jasotzen dira, nodoaren atributuetan gordez. Ordain guztietan ez dira eremu guzti horiek agertzen.

10. Universitat Politcnica de Catalunya. http://www.upc.edu

11. http://garraf.epsevg.upc.es/freeling

Zenbaitetan transferentzia lexikala ez da egin behar: preposizioak dituzten nodoak eta erroa ez diren aditz-chunketako nodoak (aditz laguntzaileak, perifrastikoak, izenordain atonoak, etab.) transferentzia estrukturalean prozesatuko dira; zifrak, datak eta orduak dituzten nodoetan, berriz, ez dago transferentziaren beharrik, sorkuntza zuzenean lematik egingo delako.

Lexikoi elebiduna eraikitzeko, *Elhuyar* hiztegia (Elhuyar, 2000), Euskalterm[12] banku terminologikotik erdi-automatikoki erauzitako hitz anitzeko terminoen zerrenda, entitateen zerrenda elebidunak eta eskuz kodetu diren kategoria itxien hiztegia erabili dira. Iturri horietako informazio guztiarekin egokitze-prozesu bat pasa ondoren, *Matxin* sistemaren lexikoi elebiduna sortu da. Lexikoia informazio semantikoarekin aberasteko, prozesu erdi-automatiko baten bidez ezaugarri semantikoak etiketatu ditugu.

Lexikoiaren sekzio nagusiak kategoria irekiak biltzen ditu (62.000 sarrera) eta kategoria itxien sekzioan determinanteak, izenordainak eta loturazko elementuak aurkitzen dira (480 sarrera). Sarrera horietatik 14.000 hitz anitzekoak dira. Garrantzi handikoa izan da hitz anitzeko unitateak jasotzeko egindako lana, lokuzio terminologikoak ondo itzultzeak erronka handia suposatzen duelako itzulpen-automatikorako. *Matxin* sistemak lexikoian dauden hitz anitzeko unitate lexikalak analisi-fasean identifikatzen ditu, transferentzia lexikalean hitz bakarreko terminoen modura bilatzeko.

Anbiguotasun lexikalak ebazteko sistemak estrategia sinple bezain eraginkorra erabiltzen du: lexikoi elebiduneko lehenengo adieraren lehenengo ordaina hautatzen da, ordain horrek gehienetan itzulpen egokia ematen duelako[13]. Estrategia honen emaitzak onak badira ere, itzulpenaren hautapen lexikalerako desanbiguazio teknikak aztertzeari ekin diogu.

## 2. Transferentzia estrukturala

Transferentzia estrukturalak abiapuntu-hizkuntzatik datorren egitura xede-hizkuntzarako egokia bihurtzen du. Tipologikoki urrun dauden hizkuntzen egituren artean, euskararen eta inguruko hizkuntzen artean bezala, desberdintasun sintaktiko handiak ematen dira, prozesua zailduz.

Euskara hizkuntza eranskaria da, postposizioak erabiltzen dituena; espainierazko zenbait elementuren itzulpenean (artikuluak, menpeko konjuntzioak, preposizioak eta funtzio sintaktikoak) euskarazko postposizioak erabiltzen dira, zenbaitetan osagai lexikal gisa desagertuz. Espainierazko aditz-kateak eta euskarazko beren itzulpenak oso desberdinak dira elementuei eta ordenari dagokienez; perifrasien itzulpenetan ematen dira desberdintasun handienak.

12. http://www1.euskadi.neteuskalterm

13. Prozesaketa automatiko batez zenbait kasutan ordainen ordena aldatu da: sustraikide bat (*cognate*) edo adiera desberdinetan errepikatzen den ordain bat aurkitzen denean.

Espainierazko egituratik euskarazko egiturara transferentzia burutzeko prozesua hainbat modulutan banatu dugu, ataza bakoitza maila batean egiten duela lan, chunk barruko nodoen artekoan edo esaldi barruko chunken artekoan:

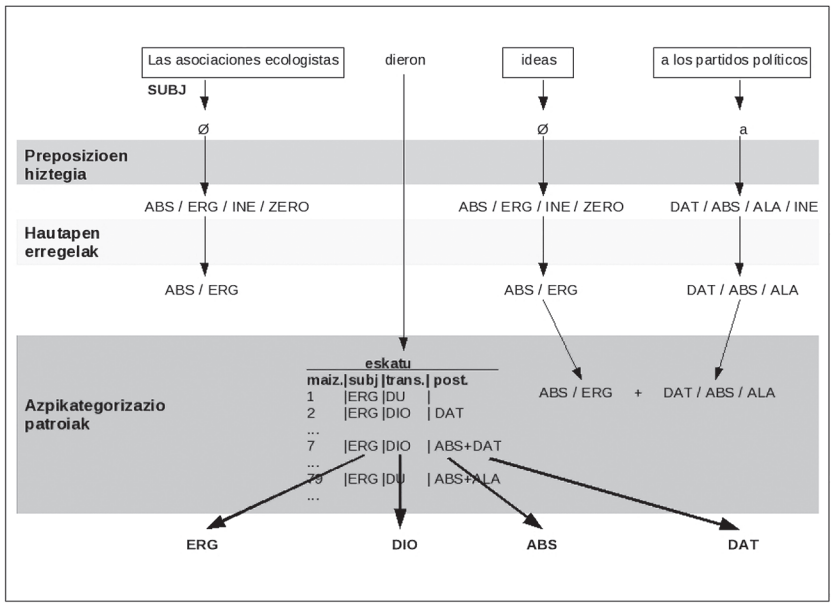
1. Chunk barruko eragiketak. Nodoetako hainbat atributu chunkera kopiatzen dira, hurrengo moduluetako eragiketetan chunkean izatea beharrezkoa izango delako. Euskarara itzultzean informazio lexikala galdu duten nodoak ezabatzen dira.
2. Preposizio eta funtzio sintaktikoen transferentzia. Prozesaketa chunk mailakoa da eta beharrezkoa den nodoetako informazioa chunketara mugitu da aurreko urratsean (ikus 4.2.1. atala).
3. Chunken arteko eragiketak. Hainbat atributu chunketik chunkera kopiatzen dira eta nodorik ez duten chunkak ezabatzen dira.
4 Aditz-kateen transferentzia. Aurreko urratsean mugitutako informazioa erabiltzen da. Abiapuntu-hizkuntzako egiturako nodoak ezabatu egiten dira eta aditz-katearen transferentziaren emaitzarekin egitura berria sortu (ikus 2.2 atala).
5. Egokitzapen-eragiketak.

### 2.1 Preposizio eta funtzio sintaktikoen transferentzia

Preposizioen itzulpena ataza zaila eta garrantzitsua da IAko sistema batean, ezin delako modu sistematikoan egin. *Matxin* sisteman aditzen modifikatzaile diren preposizio eta funtzio sintaktikoen transferentzia burutzeko hiru urrats ematen dira:

1. Preposizioen hiztegia eta hautapen-erregelak.
   Preposizioaren itzulpena hautatzeko kanpo- eta barne-argumentuetako informazio lexiko, sintaktiko eta semantikoa erabiltzen da, preposizioen hiztegiko hautapen-erregelak aplikatuz.
   Eskuz eraiki dugun hiztegi honetan preposizio bakoitzarekin (funtzio sintaktikoa adierazten duen preposizio hutsaz gain, 18 preposizio sinple eta 333 konposatu) bere itzulpena izan daitezkeen postposizio posibleak kodetu dira (guztira 462) eta, ahal denean, baita postposizio horri lotutako hautapen-erregela ere (guztia 89 erregela). Hautapen-erregelek postposizioak hautatu edo baztertzen dituzte. Postposizio bat edo gehiago hautatu badira, hurrengo urratsera pasako dira, eta ez bada postposiziorik hautatu baztertu ez direnak pasako dira.

2. Azpikategorizazio-patroiak.
   Sistemak aditz-kate baten azpiko preposizio-sintagmen preposizio guztiak batera ebazten saiatzen da.
   *Euskaldunon Egunkaria*ren testuekin sortutako corpus batetik erauzitako azpikategorizazio-patroiak (Aldezabal *et al.*, 2002) erabiltzen ditugu. Patroietan datu hauek ematen dira: aditzaren lema, subjektuaren postposizioa, beste osagaien posposizioen zerrenda

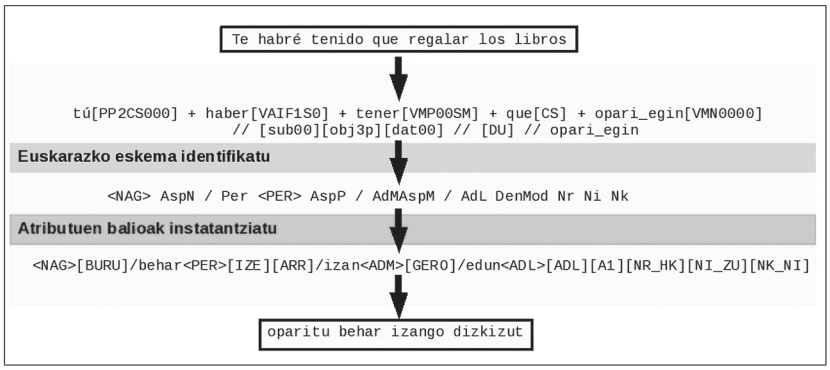



Irudia 8: Preposizioen transferentziaren adibidea

eta konbinazio horri dagokion aditzaren iragankortasuna eta patroi horren maiztasuna. Posposizioak desanbiguatzeko, subjektua alde batera utzi eta aditzaren menpe dauden beste chunketako aurreko urratsean emandako postposizio posibleen konbinazioak aztertzen ditugu, konbinazio horietako batekin bat datorren maiztasun handieneko azpikategorizazio-patroia bilatzeko. Guztiz bat datorren patroirik ez badago, elementu komun gehien dituen patroia jasoko da eta desanbiguatu gabe gelditzen diren postposizioak hurrengo urratsean ebatziko dira.

3. Hiztegiko ordena.

Adibidez (ikus 8. irudia), *Las asociaciones ecologistas dieron ideas a los partidos políticos* esaldia itzultzeko subjektuaren postposizioaz gain beste bi chunketako postposizioak desanbiguatu behar dira: preposizio hutsa duen bat eta a preposizioa duen bestea. Aurreneko urratsean preposizioen hiztegiak, hautapen-erregelak erabiliz, preposizio hutsa duen chunkarentzat absolutiboa (*ABS*) eta ergatiboa (*ERG*) eman ditu, eta a preposizioa duenarentzat datiboa (*DAT*), absolutiboa (*ABS*), eta alatiboa (*ALA*). Postposizio horien konbinazio posibleentzat eman aditzaren patroiekin bat datozenak (*ABS-DAT* eta *ABS/ALA*) bilatzen dira, maiztasun handienekoa (*ABSDAT*) hautatuz. Beraz, *DIO/ERG/ABS-DAT* patroia jasotzen da, subjektuari ergatiboa (*ERG*) emanez, aditzari *DIO* iragankortasun-informazioa eta aditzaren beste osagaiei,




Irudia 9: Aditz-kateen transferentziaren adibidea

preposizio hutsa zuenari absolutiboa (*ABS*) eta a preposizioa zuenari datiboa (*DAT*). Sistemak emandako itzulpena hauxe da: *Elkarte ekologistek ideiak eman zizkieten alderdi politikoei.*

Izen-sintagmak modifikatzen dituzten preposizioak modu sinpleagoan ebazten dira: hautapen-erregelak erabili ondoren, desanbiguatu ez direnentzako hiztegiko ordenak ebazten du.

### 2.2 Aditz-kateen transferentzia

Aditz-kateen transferentzia egoera finituetako transduktoreen bidez egiten dugu. Sortu dugun gramatikaren erregelek espainierazko aditz-kate bat jaso, zenbait transformazio burutu eta euskarazko aditz-katea sortzen dute.

Gramatikarako sarrera izango den karaktere-kateak ondoko informazioa dauka: espainierazko aditz-katearen nodo guztien informazio morfologikoa; subjektuaren, objektu zuzenaren eta zeharkako objektuaren informazio morfologikoa; eta euskarazko aditz nagusiaren iragankortasuna.

Gramatikako erregelak hiru multzotan antolatu ditugu, aditz-kateak itzultzeko definitu diren hiru urratsekin lotuta daudenak:

1. Espainierazko aditz-katearen motari dagokion euskarazko aditz-katearen eskema identifikatu. Badira hori egiteko sei erregela, bakoitza espainierazko ondoko aditz-kate mota bati dagokiona: aditz ez-jokatuak, aditz arruntak ez-perifrastikoak, eta perifrastikoen artean definitu diren lau multzoak.
2. Euskarazko eskemako atributuen balioak instantziatu.
3. Soberazko informazioa ezabatu.

Irteerako elementu bakoitzarekin nodo bat eraiki beharko da, irteeran ematen den lehenengoa erroa izango dela, eta besteak bere menpeko nodoak. Nodo bakoitzerako ematen da sor-

kuntza sintaktikoan ordena erabakitzeko informazioa eta sorkuntza morfologikoa burutzeko beharrezkoa dena. Aditzkateen transferentziaren adibide bat 9. irudian ikus daiteke.

### *Sorkuntza*

Sorkuntzaren helburua transferentzia-fasean lortutako egituratik xede-hizkuntzako testua ematea da. Sorkuntza bi mailatan burutzen da: sintaktikoa eta morfologikoa.

#### 1. Sorkuntza sintaktikoa

Sorkuntza sintaktikoak hitzak eta osagaiak xede-hizkuntzako sekuentzia zuzen batean ordenatzen ditu. Ordenazioa burutzeko transferentziatik datorren egitura sakoneko zuhaitza erabiltzen da, egitura hori oso egokia baita ordenazioa bi mailatan egiteko: chunk bakoitzeko nodoak ordenatzen dira alde batetik, eta esaldiko chunkak bestetik.

Chunk barruko sorkuntza sintaktikoa burutzen duen moduluak nodoak ordenatzen ditu chunk barruan eta postposizio-informazioa duten chunketan informazio hori chunkaren azkeneko nodoan kopiatuko du, nodo horren sorkuntza morfologikoan erabili beharko da eta. Euskaraz sintagma bakoitzaren elementuak modu jakin eta zurrun batean ordenatzen dira; beraz, aurrekotasun-erregela desberdin bat kodetu da chunk mota bakoitzerako: baiezko aditz-kateak, ezezko aditz-kateak eta bestelako chunkak (izen-sintagmak eta adjektibo-sintagmak).

Chunken arteko ordenaketa bi urratsetan ebazten dugu:

1. Guraso-ume chunk-bikote bakoitzerako ordena erlatiboa ebatzi. Erregeletan hartzen dira kontuan umearen mota, informazio sintaktikoa eta fokoa, eta gurasoaren mota. Euskarazko perpauseko chunken arteko ordena oso librea bada ere, badira zenbait gomendio chunkak ordenatzeko eta, ordenazio posible bat baino gehiago dagoenean, orokorrean egokiena deritzoguna kodetu dugu erregeletan.
2. Chunk guztien ordena absolutua ebatzi. Chunken sekuentzia ordenatua kalkulatzen da guraso-ume chunk-bikoteen ordena erlatiboaren informazioa erabiliz. Hortik chunk bakoitzari sekuentzia horretan duen ordena absolutua esleitzen zaio.

#### 2. Sorkuntza morfologikoa

Sorkuntza morfologikoaren helburua xede-hizkuntzako hitzen formak sortzea da, horretarako elementu lexikal etiketatuak interpretatu behar dituela.

Soilik postposizio-informazioa duten hitzak prozesatuko dira sortzaile morfologikoarekin, besteetan hitzaren forma lemaren forma bera izango da eta. Aditz-kateetan nodo guztiek sorkuntza egiteko informazioa izango dute eta beste chunketan soilik chunkeko azken elementua. Zenbakiekin, datekin eta hitz ezezagunekin sorkuntza morfologiko berezia egiten da.

Prozesadore morfologikoak postposizio-informazioaz gain lemaren informazio morfologikoa behar du. Lexikoi elebidunetik jasotako lemak osaketa morfologikorik baldin badu, osaketa horren informazioa erabiliko da; bestela, lema eta *part-of-speech* informazioa nahikoa izango da.

Sorkuntza morfologikoa burutzeko IXA taldean garatutako *Morfeus* euskararako prozesadore morfologikoa (Alegria eta Urkia, 2002) berrerabili dugu, 60.000 sarrera inguru dituen EDBL euskararen datu-base lexikalean oinarritzen dena.

| Itzulpen egokiak | |
|---|---|
| Le llevé el pan a mi hermano a casa | Ogia eraman nion nire anaiari etxera |
| Viene en coche y vive en esta ciudad | Automobilaz dator eta hiri honetan bizi da |
| Los políticos dicen que demos tiempo al tiempo | Politikariek esaten dute pazientzia izan dezagula |
| Los aviones volaron sobre la muchedumbre | Hegazkinek jendetzaren gainetik hegan egin zuten |
| El libro está sobre la mesa | Liburua mahaiaren gainean dago |
| Itzulpen traketsak | |
| Cuatro nuevas sucursales de Correos se abrirán en la capital | Correos-en 4 sukurtsal berri kapitalean irekiko dira |
| El hospital tendrá 48 nuevas habitaciones individuales en 2009 | Ospitaleak 48 banako gela berri izango du 2009tan |
| Fue entonces cuando escuchó la explosión que se produjo en el primer piso | Orduan izan zen leherketa entzun zuenean eragin zen 1 pisuan |
| Mientras en la Unión Europea la edad media de independizarse son 22 años, en España supera los 26 | Europar Batasunean Erdi Aroa banandu bere burua izatera 22 urtetan izan, Espainian 26 gainditzen du |

Taula 1: Itzulpen-adibideak

## *Ebaluazioa*

*Matxin* sistemak itzulpen egokiak ematen ditu kasu askotan, batez ere esaldiak sinpleak direnean, baina bestetan lortutako itzulpenak nahiko traketsak dira (ikus 1. taula).

Espainieratik euskarara itzultzea ataza konplexua da. Emandako itzulpenak oinarrizko ulermenerako baliagarriak izan daitezke eta, beraz, asimilaziorako sistema sendo bat eraikitzeko bidean, *Matxin* sistemak etorkizun oparoa du.

## 1. Ebaluazioaren emaitzak

Azken hamarkadan IArako ebaluazio-neurririk erabiliena bilakatu den Bleu (Papineni *et al.*, 2002) metrikaren inguruan zalantza ugari sortu izan dira (Callison-Burch *et al.*, 2006; Koehn eta Monz, 2006): ez du IAko sistemen itzulpenaren kalitate absolutua neurtzen, ezta erabiltzaile batentzat itzulpenak zenbateraino diren baliagarriak erakusten. Gainera estrategia desberdinetako sistemak konparatzeko ere ez du balio.

Gure sistema ebaluatzeko *HTER* (*Human-targeted Translation Edit Rate*) neurria (Snover *et al.*, 2006; Przybocki *et al.*, 2006) aukeratu dugu. *HTER* kalkulatzeko giza editore batek IAko sistema baten itzulpenean eginbeharreko moldaketak burutzen ditu, editatutako bertsioak abiapuntuko testuaren esanahi osoa izan dezan idazkera ulergarrian. Moldaketa posibleak dira banakako hitzen txertatzea, ezabaketa, ordezpena eta hitz multzoen mugitzea. Ondoren edizio kopurua zati moldatutako itzulpenaren hitzen kopurua kalkulatzen da.

Ebaluaziorako neurri intuitibo honek sistemaren itzulpenen kalitatea modu errealistan neurtzen du, itzulpenak zenbateraino diren baliagarriak erakutsiz.

Ebaluazioa burutzeko testuak bi corpus desberdinetatik jaso ditugu: Eitb, hizkuntza orokorreko kazetaritza-corpusa, eta *Consumer* (Alcázar, 2006), kontsumoaren arlokoa. Bi corpus horietako bakoitzetik 5 eta 25 bitarteko hitz kopurua duten 50 esaldi aukeratu dira ausaz.

Gainera, *Matxin*en emaitzak konparatu ahal izateko, espainieratik euskarara itzultzen duen *Matrex* corpusetan oinarritutako sistema baldintza berdinetan ebaluatu dugu. Dublinen garatutako Matrex sistema euskarara itzultzeko egokitu izan da, IXA taldearen euskararako hainbat tresna erabiliz (Labaka *et al.*, 2007), eta Consumer aldizkariko corpuseko 50.000 esaldiekin entrenatu da.

Ebaluazioaren emaitzak 2. taulan ikus daitezke. HTER balio txikiagoek kalitate hobeagoa adierazten dute. *Matxin* sistemarako emaitzen batazbestekoa 42koa da, hau da, 100 tokenetatik 42 edizio burutu behar izan dira. *Eitb* corpus periodistikoko adibideekin emaitzak hobeagoak dira, *Consumer* corpuseko esaldietan domeinu zehatz bateko egitura sintaktiko eta terminologia berezitu gehiago agertzen delako.

*Matxin* eta *Matrex* sistemen emaitzak konparatzen baditugu, ikus dezakegu *Consumer* corpuseko esaldietan, *Matxin*en irteera *Matrex*ena baino hobeagoa dela (43.60 vs. 57.97), eta *Matrex* entrenatu ez den *Eitb*ko corpusekoetan diferentzia oraindik askoz handiagoa dela (40.41

| | *HTER* | |
|---|---|---|
| | *Matxin* | *Matrex* |
| *Eitb* | 40.41 | 71.87 |
| *Consumer* | 43.60 | 57.97 |

Taula 2: Ebaluazioaren emaitzak

vs. 71.87). *Matrex* sistema hobe daiteke corpus handiagoekin entrenatuz eta hainbat doikuntza eginez, baina ikustekoa da zenbateko hobekuntza lor dezakeen corpusetan oinarritutako sistema honek.

## 2. Erroreen analisia

Ondokoak dira erroreen iturburu nagusiak, sortzen duten errore-kopuruaren arabera ordenatuta:

1. Analisia (erroreen %43). Erroreak ematen dira nagusiki mendekotasun-analizatzailean (batez ere loturazko elementuekin eta *PP-attachment* delakoarekin), etiketatzaile morfosintaktikoan eta funtzio sintaktikoen esleipenean. Adibidez, ondoko esaldian espainierazko analizatzaileak *nota* izena *notar* aditzaren formatzat hartu du:

| | |
|---|---|
| Qué nota tiene? | Du zerk nabari du? |

2. Hautapen lexikala (%14).

| | |
|---|---|
| Médicos especialistas asesorarán a los de cabecera | Trebe medikuek oheburuak aholkatuko dituzte |

3. Preposizio eta funtzio sintaktikoen itzulpena (%8).

| | |
|---|---|
| en el colegio de Rojales | Rojales-en ikastetxean |

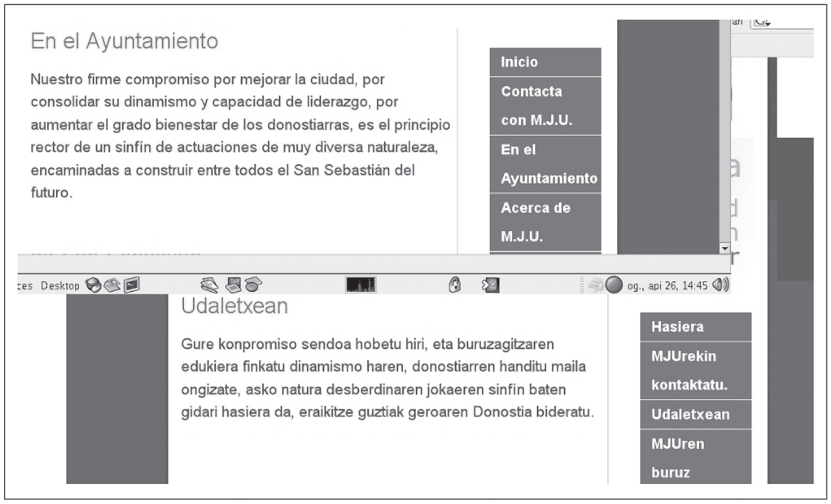
4. Hitz anitzeko terminoak (%6).

| | |
|---|---|
| casos de siniestralidad | ezbehar-kopuruaren kasuak |

5. Beste errore batzuk, neurri txikiagoan, gertatzen dira: hitzen ordenan, aditzen transferentzian hiztegiarekin (hainbat sarrera ez egotea edo ordain okerrak ematea), sorkuntza morfologikoan, aditzarekin komunztadura okerrak, zenbaki eta laburduren itzulpenetan eta formatu-arazoak (kakotxak, maiuskulak, etab.).

## 3. Erabilera okerrak

Garrantzitsua da azpimarratzea *Matxin* ez dela zabalkunderako sistema bat, hau da, ezin direla argitaratzeko erabili automatikoki lortzen diren itzulpenak, ez badira lehenago posteditatu eta zuzendu. Hala ere, zenbait kasutan erabilera okerrak egiten ari dira Interneten erabilgarri dagoen *Matxin* sistemak emandako itzulpenekin, besteak beste, euskarari errespetu gutxi erakutsiz. 10. irudian ikus dezakegu alderdi politiko batek hauteskunde kanpaina baterako Interneten argitaratutako web-orria, eta 11. irudian merkatalgune batean agertutako kartela. Bi kasuetan une hartako *Matxin* sistemaren bertsio publikoak espainierazko testurako emandako itzulpena erabili zen, inolako zuzenketarik egin gabe.




Irudia 10: Itzulpen automatikoaren erabilera okerrak (I)


Irudia 11: Itzulpen automatikoaren erabilera okerrak (II)

### *Ondorioak eta etorkizunerako lanak*

Gure helburuak bete ditugu: erregeletan oinarritutako itzulpen-sistema diseinatu eta inplementatzeaz gain, *Matxin* espainiera-euskara sistema publikoki erabilgarri dago eta kode irekiko software libre bezala banatzen da.

Argi dago euskararako itzulpen automatikoko sistemen behar handia dagoela: *Matxin* sistema *Opentrad* proiektuaren web-orrian erabilgarri jarri zenetik erabilera oso handia izan du, egunero 4.000 itzulpen inguru egiten direla.

Sistemak bere lana abiadura onargarrian betetzen du (300 hitz/segundoko) eta atazaren konplexutasuna kontuan hartuta, sistemaren itzulpenetan lortutako kalitatea positiboki baloratzen dugu, oinarrizko ulermenerako balio duelako, zabalkunde-sistema batek izan behar duen zuzentasunetik urrun badago ere. Gure sistemaren eraikuntzak euskara bezalako hizkuntzentzat erregeletan oinarritutako estrategiaren ahalmena frogatzen du eta estrategia horren zailtasunak eta mugak identifikatzen lagundu digu.

Berriki, *Matxin* itzultzailea *AnHitz* (Arrieta *et al.*, 2008) proiektuan integratu da, euskaraz hitz egiten duen zientzia-aditu birtual interaktiboa eraikitzeko. Prototipo horren lehenengo ebaluazioan *Matxin*ek emandako itzulpenen kalitateaz pozik egon gaitezke: itzulpenen %69a ulertzeko modukoa da.

*Matxin* sistemaren itzulpenen kalitatea hobetzeko, sistemako moduluetan eta datu linguistikoetan erroreen arazketa eta zuzenketa egiten ari gara, eta nahitaezkoa da espainerarako analizatzailea hobetzea. Gure sistemak dituen mugak gainditzeko beste hainbat teknika eta estrategia aztertzen eta erabiltzen hasiak gara: hitzen adiera-desanbiguaziorako tekniken azterketa hautapen-lexikalerako, domeinu zehatzetarako egokitzapenak (adibidez, lan-hitzarmenak eta telefoniako esku-liburuak), corpusetan oinarritutako estrategien ikerketa eta *Matxin* estrategia horiekin konbinatzea edo hibridatzea. Gainera, ingelesetik euskarara itzultzen duen prototipoak laster ikusiko du argia.

Badira ere abiatu dugun lanaren jarraipen zuzena izango diren hainbat ataza: postediziorako interfazearen diseinua eta inplementazioa, lexikoia aberasteko tresnen inplementazioa, beste norabiderako sistema bat eraikitzeko *Matxin* sistema berrerabiltzeko aukeren azterketa, eta *Matxin* sistemaren ahalmena frogatzeko, euskara barne hartzen ez duen sistema baten inplementazioa (adibidez, espainieratik quechuara).

Egindako lana aitzindaria da, bai euskararekin lan egiten duen lehenengo IAko sistema erabilgarria delako eta bai software librean eraikitako lehenengo IAko sistemetako bat delako.

Umiltasunez aitortzen dugu guk egindako lana hasiera baino ez dela; harrotasunez onartzen dugu gure ekarpenek bide asko zabaltzen dituztela. Itzulpen automatikoa eta euskararekin lan egiten eta egingo dutenentzat guk egindako lana erabilgarria izatea espero dugu.

## ERREFERENTZIAK


Alcázar A. Towards linguistically searchable text. In *Proceedings of BIDE 2005*, Deusto. Bilbao, 2006.

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., and Sarasola K. Learning argument/adjunct distinction for basque. In *ACL'2002 SigLex Workshop on Unsupervised Lexical Acquisition*, 2002.

Alegria I. and Urkia M. *Morfologia konputazionala. Euskararen morfologiaren deskribapena.* UEU, 2002. ISBN 84-8438-034-3.

Arrieta K., de Ilarraza A.D., Hernez I., Iturraspe U., Leturia I., Navas E., and Sarasola K. Anhitz, development and integration of language, speech and visual technologies for basque. In *Second International Symposium on Universal Communication*, JAPAN, 2008.

Atserias J., Casas B., Comelles E., González M., Padró L., and Padró M. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006.

Atserias J., Comelles E., and Mayor A. TXALA, un analizador libre de dependencias para el castellano. In *Actas del XXI Congreso de la SEPLN XXI. Demo session*, pages 455–456, 2005.

Callison-Burch C., Osborne M., and Koehn P. Re-evaluating the role of BLEU in *Machine Translation research. In Proceedings of EACL-2006*, 2006.

Civit M. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español.* PhD thesis, Universidad de Barcelona, 2003.

Corbí-Bellot A.M., Forcada M.L., Ortiz-Rojas S., Pérez-Ortiz J.A., RamírezSánchez G., Sánchez-Martínez F., Alegria I., Mayor A., and Sarasola K. An open-source shallow-transfer Machine Translation engine for the romance languages of Spain. In *Proceedings of the EAMT2005*. Poster session, Budapest, Hungary, 2005.

Elhuyar. *Elhuyar Hiztegia*. Elhuyar Hizkuntz Zerbitzuak, 2000. ISBN 8495338-08-4.

Forcada M.L., Bonev B.I., Rojas S.O., Ortiz J.A.P., Sanchez G.R., Martínez F.S., and Rosell M.G. Documentación del sistema de código abierto Opentrad Apertium de traducción automática de transferencia sintáctica superficial. Technical report, Departament de Llenguatges i Sistemes Informatics. Universitat d'Alacant, 2006.

Koehn P. and Monz C. Manual and automatic evaluation of Machine Translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics.

Labaka G., Stroppa N., Way A., and Sarasola K. Comparing rule-based and data-driven approaches to spanish-to-basque machine translation. In *Proceedings of the MT-Summit XI*, Copenhagen, 2007.

Papineni K., Roukos S., Ward T., and Zhu W. BLEU: a method for automatic evaluation of Machine Translation. In P*roceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

Przybocki M., Sanders G., and Le A. Edit distance: a metric for Machine Translation evaluation. In *Proceedings of the LREC-2006: Fifth International Conference on Language Resources and Evaluation*, pages 2038–2043, Genoa, Italy, 2006.

Snover M., Dorr B., Schwartz R., Micciulla L., and Makhoul J. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, 2006, 2006.

**_Matxin_, el primer sistema de traducción automática que traduce a euskera**

En este artículo presentamos *Matxin*, el primer sistema de traducción automática que traduce a euskera. Es un sistema basado en reglas, que sigue el modelo tradicional de transferencia. La reutilización de diferentes herramientas y recursos lingüísticos de amplia cobertura ha hecho necesaria la estandarización de los formatos para garantizar la interoperatividad. Esto, a su vez, posibilitará en el futuro extender el sistema a otros idiomas e integrar nuevos módulos. El prototipo Matxin 1.0, que traduce de español a euskera, puede usarse en Internet (www.opentrad.org) y se distribuye como software de código abierto (www.matxin.sourceforge.net). Las traducciones realizadas por el sistema pueden servir para entender el texto original, pero su calidad no es aún suficiente para que sean utilizables para su publicación. El sistema está siendo adaptado para traducir de inglés a euskera.

**_Matxin_, le premier système de traduction automatique qui traduit en euskara**

Dans cet article nous présentons *Matxin*, le premier système de traduction automatique qui traduit en euskara. C'est un système fondé sur des règles et qui suit le modèle traditionnel de transfert. La réutilisation des différents instruments et des ressources linguistiques amplement répandues a rendu nécessaire la standardisation des formats pour garantir l'interopérativité. À l'avenir, cela rendra possible l'extension du système à d'autres langues et l'intégration de nouveaux modules. Le prototype Matxin 1.0, qui traduit de l'espagnol en basque peut être utilisé sur Internet (www.opentrad.org) et il est distribué comme software de code ouvert (www.matxin.sourceforge.net). Les traductions réalisées par le système peuvent servir à comprendre le texte original, mais leur qualité n'est pas encore suffisante pour qu'elles soient publiées. En ce moment, on est en train d'adapter le système pour traduire de l'anglais en euskara.

**_Matxin_, the First Automatic Translation System that Translates into Basque**

In this article, we introduce *Matxin*, the first automatic translation system that translates into Basque. It is a system based on rules, and that follows the traditional transfer model. The reuse of different tools and extensive linguistic resources has made it necessary to standardize formats in order to guarantee interoperativity. This, in turn, allows the future extension of the system to other languages and the integration of new modules. The Matxin 1.0 prototype, which translates from Spanish into Basque, can be used on internet (www.opentrad.org) and is distributed as open code software (www.matxin.sourceforge.net). The translations created by the system can be used to clarify the original text, but are not yet of sufficient quality for publication. The system is currently being adapted to translate from English into Basque.