

Matxin

Erregeletan oinarritutako itzulpen automatikoko
sistema baten eraikuntza
estaldura handiko baliabide linguistikoak berrerabiliz

Aingeru Mayor Martinez

Lengoaia eta Sistema Informatikoak Saila
Euskal Herriko Unibertsitatea

2007ko azaroaren 27a

Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena
- 4 Baliabide linguistikoaren berrerabilpena
- 5 Ebaluazioa
- 6 Ondorioak eta etorkizunerako lanak

Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena
- 4 Baliabide linguistikoaren berrerabilpena
- 5 Ebaluazioa
- 6 Ondorioak eta etorkizunerako lanak

Sarrera

Itzulpen automatikoa (IA):

Sistema informatikoak erabiliz
hizkuntza batetik beste batera
itzulpena burutzen duen prozesua

- Gure mundu globalizatuan IAK garrantzi handia hartu du
- Erabilera nagusiak:
 - Asimilazioa: oinarritzko ulermenerako
 - Zabalkundea: argitaratzeko kalitatezko itzulpenak

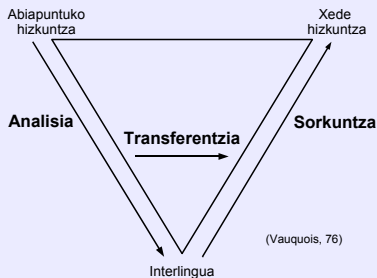
Estrategiak

- Erregeletan oinarritutako itzulpen automatikoa:
 - Ezagutza linguistikoa adituek kodetzen dute erregeletan
- Corpusetan oinarritutako itzulpen automatikoa:
 - Ezagutza corpusetatik jasotzen da
 - Bi hurbilpen:
 - Adibideetan oinarritutakoa (Nagao, 84)
 - Itzulpen automatiko estatistikoa (Brown *et al.*, 90)

Etorkizuna hibridazioaren bidetik etorriko da

Erregeletan oinarritutako itzulpen automatikoa

- Itzulpen-prozesuaren faseak:
 - Analisia
 - Transferentzia
 - Sorkuntza
- Estrategiak:
 - Itzulpen zuzena
 - Interlingua bidez
 - Transferentzian oinarrituta



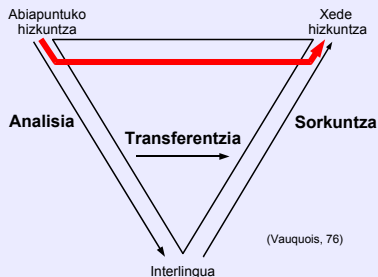
Erregeletan oinarritutako itzulpen automatikoa

- Itzulpen-prozesuaren faseak:

- Analisia
- Transferentzia
- Sorkuntza

- Estrategiak:

- Itzulpen zuzena
- Interlingua bidez
- Transferentzian oinarrituta



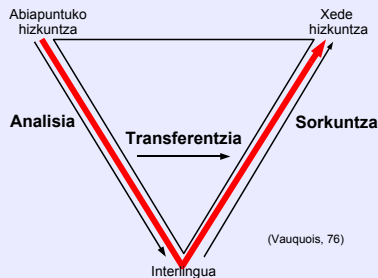
Erregeletan oinarritutako itzulpen automatikoa

- Itzulpen-prozesuaren faseak:

- Analisia
- Transferentzia
- Sorkuntza

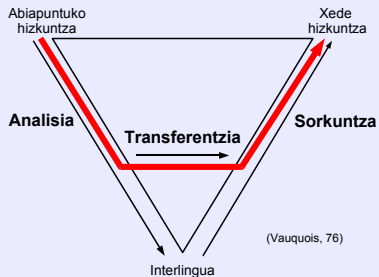
- Estrategiak:

- Itzulpen zuzena
- Interlingua bidez
- Transferentzian oinarrituta



Erregeletan oinarritutako itzulpen automatikoa

- Itzulpen-prozesuaren faseak:
 - Analisia
 - Transferentzia
 - Sorkuntza
- Estrategiak:
 - Itzulpen zuzena
 - Interlingua bidez
 - **Transferentzian oinarrituta**



Itzulpen automatikoa eta kode irekiko software librea

- IAko sistementzat kode irekiaren abantailak:
 - Komunitate ireki batek hobekuntzak egitea
 - Modulu berriak integratzea
 - Erabilera zehatzetarako egokitzea
 - Hizkuntza berriak integratzea
- IArako kode irekiko sistema erabilgarri gutxi:
 - *Apertium* (Armentano-Oller *et al.*, 07):
 - Azaleko transferentzia bidezko sistema
 - *Logos Open Source Machine Translation*:
 - *Logos* (Scott, 03) sistema komertzialaren kode-irekiko bertsioa
 - *Matxin*

Itzulpen automatikoa eta euskara

- Azkeneko urteetan zenbait ikerketa abian jarri dira:
 - Erregeletan oinarrituta
 - *Matxin*
 - *ATS-euskara* bideragarritasun-azterketa (AutomaticTrans, 03)
 - Corpusetan oinarrituta
 - Corpus elebidunen parekatze eta prozesaketa (Abaitua, 97; Casillas, 00; Casillas *et al.*, 06;)
 - Itzulpen-memoretako sistemak (Ortiz *et al.*, 03; Sanchis *et al.*, 07)
 - Eremu zehatzetarako IA (Nevado *et al.*, 04; Pérez *et al.*, 05, 06, 07)
 - *Matrex* corpusetan oinarritutakoko sistema: en→eu eta es→eu (Way *et al.*, 06; Labaka *et al.*, 07)

Itzulpen automatikoa eta euskara

- Euskararekin ondoko baldintzek corpusetan oinarritutako hurbilpenen erabilera zailtzen dute:
 - Euskararako dauden corpusen kopuru mugatua
 - Euskararen izaera eranskaria

Hipotesia eta helburua

- Hipotesia
 - Posible da euskararako asimilaziorako IAKo sistema bat sortzea
 - Euskararako dauden baliabideekin, erregeletan oinarritutako teknikak aportazio handia egin dezakete
- Helburua
 - Erregeletan oinarritutako sistema baten eraikuntza, estrategia honen:
 - ahalmena frogatzeko
 - mugak aztertzeke

Tesi honen emaitza

Erregeletan oinarritutako Matxin es→eu sistema

Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena
- 4 Baliabide linguistiko berrerabilpena
- 5 Ebaluazioa
- 6 Ondorioak eta etorkizunerako lanak

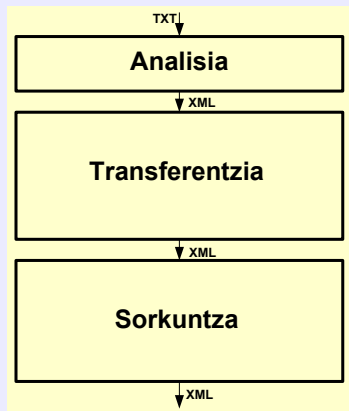
Bilakaera

- Sistemaren diseinua eta hasierako prototipoak (98-04)
 - 1. prototipoa: en→eu, izen- eta preposizio-sintagmen itzulpena
 - 2. prototipoa: es→eu, esaldi osoen itzulpena
- Opentrad proiektua (05-07)
 - Estatu espainiarreko hizkuntza nagusietarako abiadura handiko eta kode irekiko IAKo sistemen eraikuntza
 - *Apertium*
 - *Matxin*
- *Matxin* 1.0 (07)

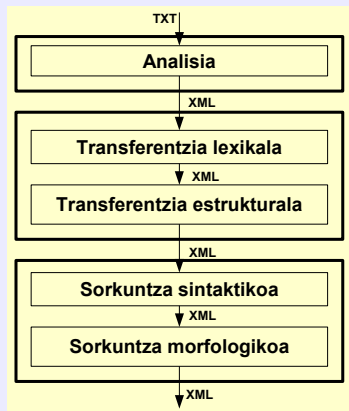
Ezaugarriak

- Transferentzia sakonean oinarrituta
- Egoera finituetako teknologiaren erabilpena
- Modulartasuna
 - Moduluen arteko banaketa ataza linguistikoek gidatua
 - Modulu elebakarrak modulu elebidunetatik ahalik eta independenteenak
 - Algoritmoak eta datuak banatuta
- Berrerabilgarritasuna
 - Elkarreragingarritasuna (*interoperability*)
 - Estandarizazioa (XML)
 - Kode irekia

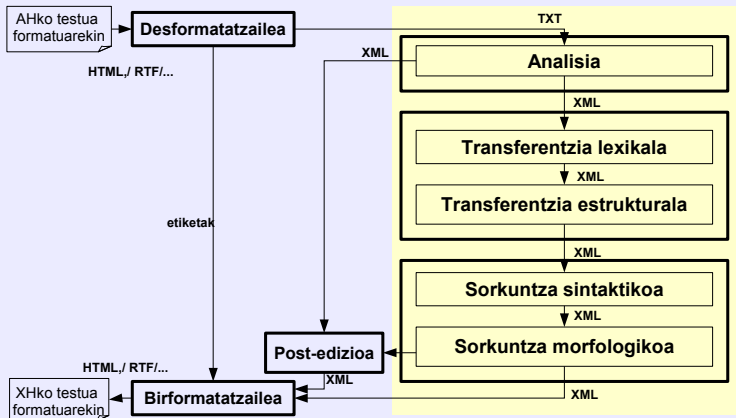
Arkitektura orokorra



Arkitektura orokorra



Arkitektura orokorra

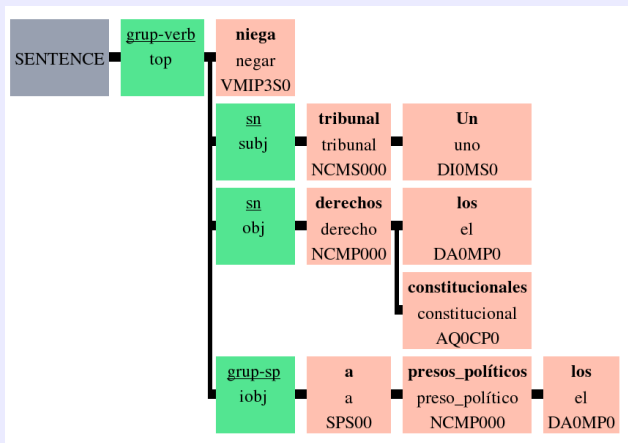


Itzulpena prozesatzeko datu-egitura

- Egitura hibridoa
 - Osagai sintaktikoen egitura
 - Hitzak esaldiko osagai nagusietan multzokatuta
 - Osagai horiek etiketatuta
 - Mendekotasun-egitura
 - Hitzen eta osagaien arteko mendekotasun-erlazioak
 - Funtzio sintaktikoak
- Hiru objektu mota:
 - Esaldia: itzulpeneko unitatea
 - Chunka: osagai bat adierazten duen sasi-sintagma
 - Nodoa: hitz bat edo hitz anitzeko unitatea
- XMLn oinarritutako formatua

Datu-egituraren formatua. Adibidea

Un tribunal niega los derechos constitucionales a los presos políticos



Datu-egituraren formatua. Adibidea

Un tribunal niega los derechos constitucionales a los presos políticos

```
<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='DIOMS0'/>
        </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='DAOMPO'/>
        <NODE ord='3' alloc='31' form='constitucionales' lem='constitucional' mi='AQOCPO'/>
      </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 politicos' lem='presopolitico' mi='NCMP000'>
        <NODE ord='2' alloc='50' form='los' lem='el' mi='DAOMPO'/>
      </NODE>
    </NODE>
  </CHUNK>
</CHUNK>
</SENTENCE>
```

Baliabide linguistiko formatua

- Baliabide linguistiko guztiak formatu estandarretan kodetuta
 - Hiztegi nagusiak: *Apertium* proiektuaren XML formatua
 - Euskarazko hiztegi morfologikoa
 - Lexikoi elebiduna
 - Beste datu linguistikoak: XML formatua
 - Preposizioen hiztegia eta hautapen-murritzpenak
 - Azpikategorizazio-patroien informazioa
 - Erregelak:
 - Datu-egitura manipulatzeko: XML eta XPath
 - Aditz-kateen transferentziarako: XFST

Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena**
- 4 Baliabide linguistiko berrerabilpena
- 5 Ebaluazioa
- 6 Ondorioak eta etorkizunerako lanak

Analisia

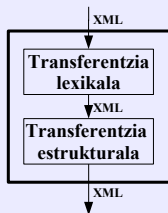
- Helburua:
 - Abiapuntu-testuaren errepresentazio abstraktua lortzea
- Analisia
 - morfologikoa
 - sintaktikoa
 - partziala
 - osoa
- Transferentzia sakona burutu ahal izateko, analisi sintaktiko osoa behar dugu

Analizatzaile baten berrerabilpena

- Espainiera analizatzeko tresna sendoak badaude, baina soilik analisi partziala ematen dute
- Estrategia:
 - Analizatzaile partzial bat erabili: *Freeling* (Atserias et al., 06)
 - Kataluniako UPC Unibertsitateko TALP taldeak garatuta
 - Kode irekia
 - Hedapena garatu:
 - Mendekotasun-erlazioak eta funtzio sintaktikoak identifikatzeko
 - Gure hedapenean oinarrituta, UPCko taldeak *FreeLing* egokitu du mendekotasun-analisisa emateko (Atserias et al., 05)

Transferentzia

- Helburua:
 - Abiapuntu-hizkuntzako testuaren adierazpide abstraktutik xede-hizkuntzako adierazpidea lortzea



Transferentzia lexikala

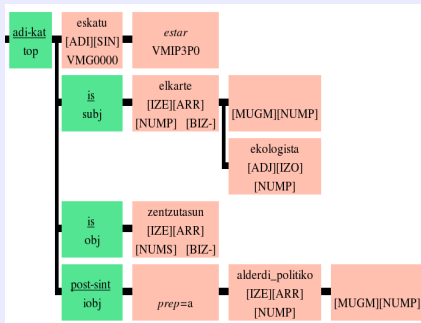
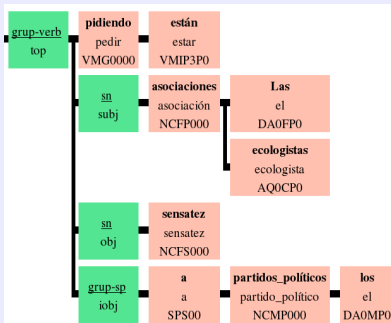
- Funtsa:
 - Hizkuntza bateko unitate lexikal bakoitza beste hizkuntzan dagokion ordainarekin ordezkatzea
- Zailtasunak:
 - Anbiguotasun lexikalak
tráfico → salerosketa, zirkulazio, trafiko
 - Estrategia: Hiztegi elebiduneko lehenengo ordaina jaso
 - Etorkizunean: Hautapen lexikalerako desanbiguazio-teknikak
 - Lokuzio terminologikoak
irse a pique → hondoratu
 - Estrategia: Analisi-fasean identifikatu eta transferentzia lexikalean hitz bakarreko terminoen modura bilatu

Transferentzia lexikalaren modulua

- Lexikoian kontsulta
 - Sarrera:
 - Abiapuntu-hizkuntzako lema eta informazio morfologikoa
 - Irteera:
 - Xede-hizkuntzako ordainaren lema, kategoria eta azpikategoria morfosintaktikoak eta beste informazio batzuk
 - Ondoko kasu hauetan ez da egiten:
 - Aditz-kateetan, aditz nagusia ez diren nodoetan
 - Preposizioak dituzten nodoetan
 - Zifrak, datak eta orduak dituzten nodoetan
- Eragiketa osagarriak

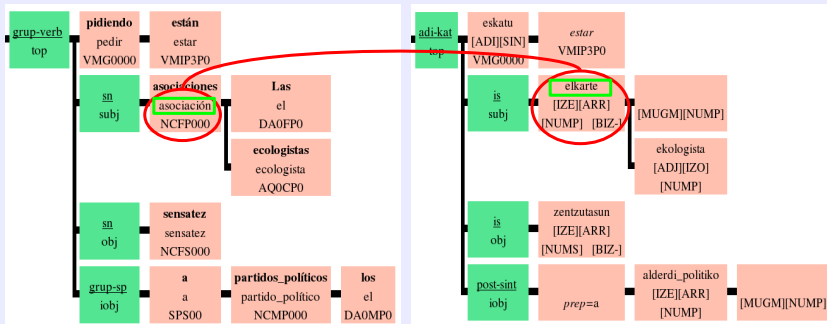
Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



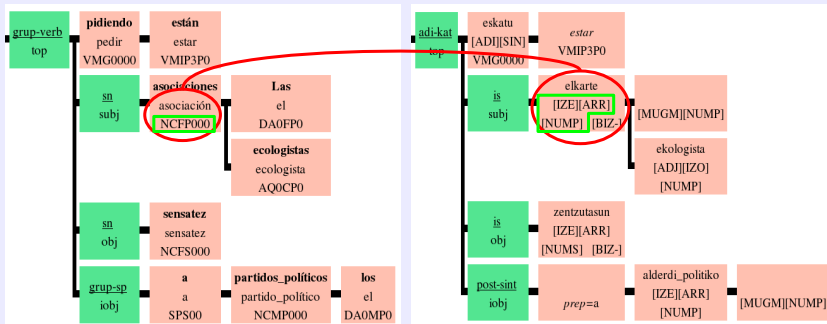
Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



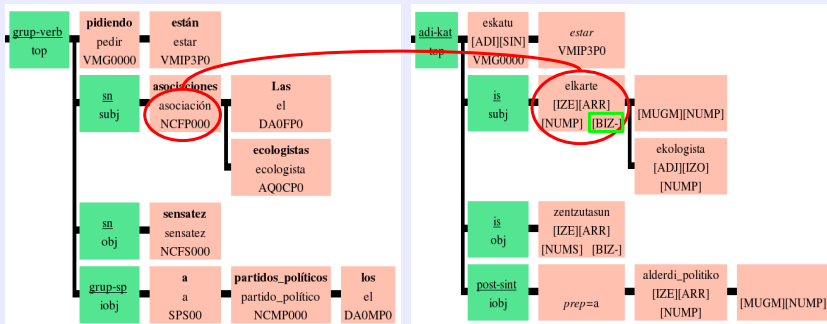
Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



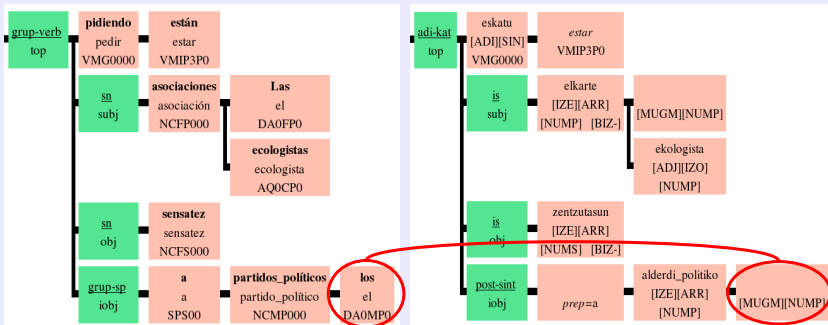
Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



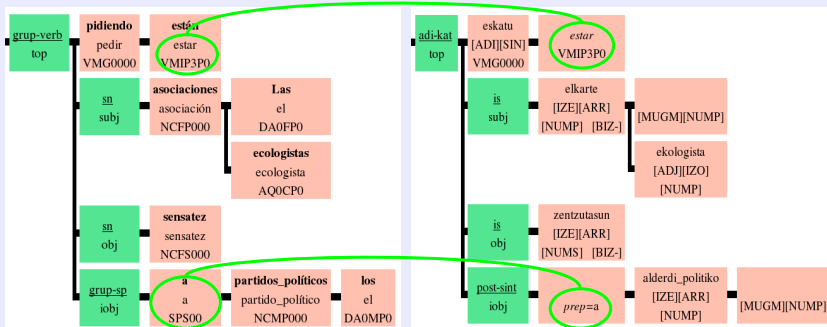
Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



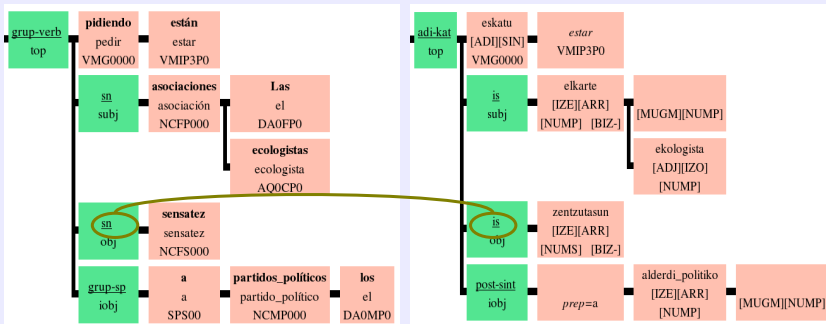
Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



Transferentzia lexikala. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



Lexikoi elebiduna

- XML formatuan kodetuta, *Apertium* proiektuaren espezifikazioari jarraituz
- Sarreraren diseinua:

Espainiera	
lem	mi

Euskara											
lem	pos	suf	loc	per	num	det	lmi	post	spost	vpost	sem

```
<e> <!aquellas>
  <p> <l> aquel <s n='mi' /> DDOFP0 </l>
    <r> haiek
      <s n='pos' /> [DET] [ERKARR]
      <s n='det' /> [MUGM]
      <s n='num' /> [NUMP]
      <s n='loc' /> [ATZ]
      <s n='lmi' /> hura [DET] [ERKARR] [NUMS]
    </r> </p>
</e>
```

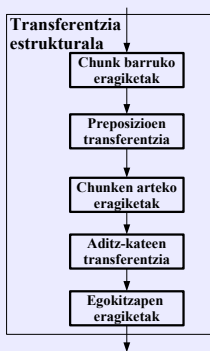
Lexikoi elebiduna

- **Kategoria irekiak**
 - Izenak, adjektiboak, aditzak eta adberbioak
 - 62.000 sarrera
 - Estaldura handiko baliabide lexikalak berrerabiliz eraikita
- **Kategoria itxiak**
 - Determinanteak, izenordainak eta loturazko elementuak
 - 480 sarrera
 - Eskuz kodetuta

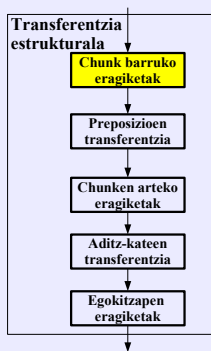
Transferentzia estrukturala

- Helburua:
 - Abiapuntu-hizkuntzatik datorren egitura xede-hizkuntzarako egokia den egitura bihurtzea
- Oso prozesu konplexua, espainiera eta euskararen artean desberdintasun sintaktiko handiak daudelako:
 - Postposizio edo atzizkitara itzultzen diren elementuak (artikuluak, menpeko konjuntzioak, preposizioak, funtzio sintaktikoak)
*el amigo dijo **que** venían **en** coche → lagunak esan zuen kotxe**z** zetozela*
 - Aditz-kateen egitura
me los han tenido que traer → ekarri behar izan dizkidate

Transferentzia estrukturalaren moduluak



Chunk barruko eragiketak

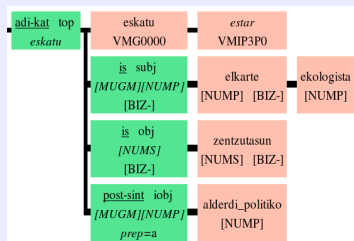
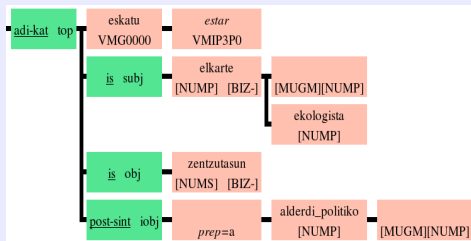


Bi eragiketa mota:

- Informazio-mugimenduak nodoetatik chunkera
 - Hurrengo moduluek burutzen duten chunk mailako prozesaketan behar delako
- Informazio lexikalik ez duten nodoen ezabaketa

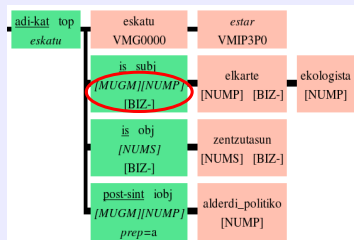
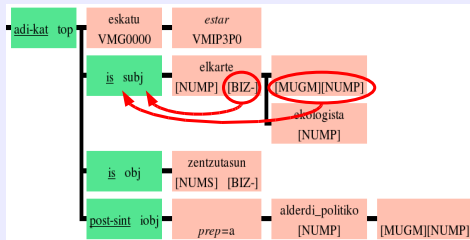
Chunk barruko eragiketak. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



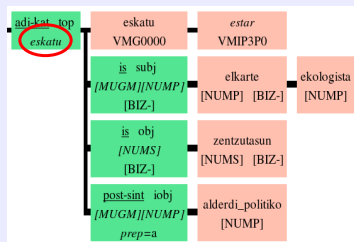
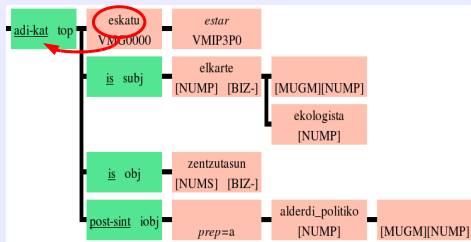
Chunk barruko eragiketak. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



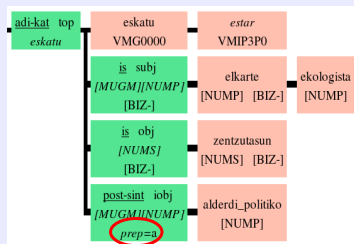
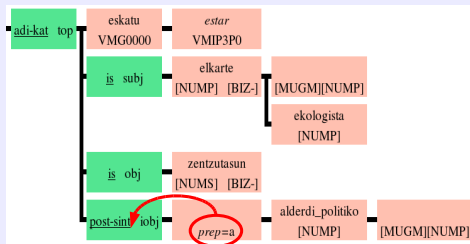
Chunk barruko eragiketak. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



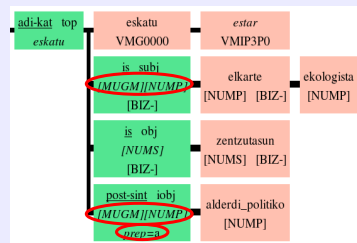
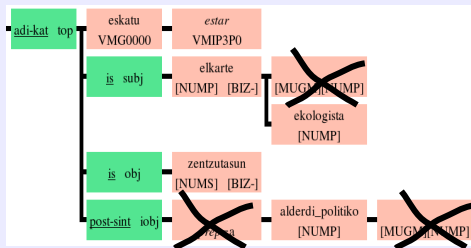
Chunk barruko eragiketak. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos

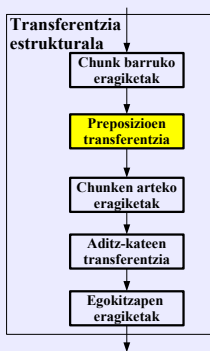


Chunk barruko eragiketak. Adibidea

Las asociaciones ecologistas están pidiendo sensatez a los partidos políticos



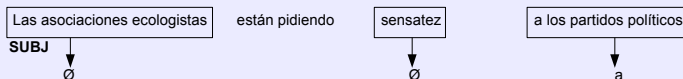
Preposizio eta funtzio sintaktikoen transferentzia



Estrategiak:

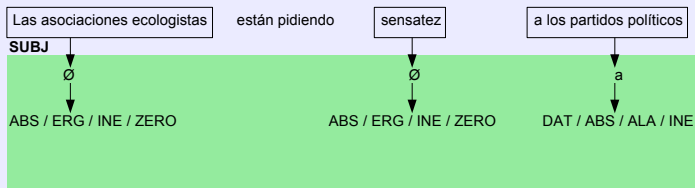
- 1 Hautapen-erregelak dituen preposizioen hiztegia
 - Eskuz kodetuta
 - Erregelek testuinguruko informazioa erabiltzen dute
- 2 Azpikategorizazio-patroien estrategia
 - Aditzaren azpiko postposizio guztiak batera desanbigutzen saiatzen da
 - Corpus batetik erauzitako (Aldezabal *et al.*, 02) azpikategorizazio-patroiak
- 3 Preposizioen hiztegiko ordainen ordena

Preposizio eta funtzio sintaktikoen transferentzia. Adibidea



Preposizio eta funtzio sintaktikoen transferentzia. Adibidea

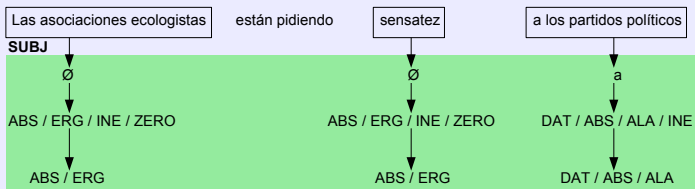
Preposizioen
hiztegia



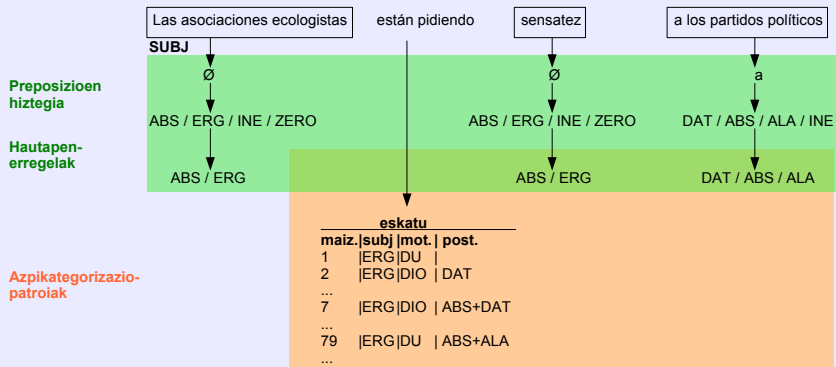
Preposizio eta funtzio sintaktikoen transferentzia. Adibidea

Preposizioen
hiztegia

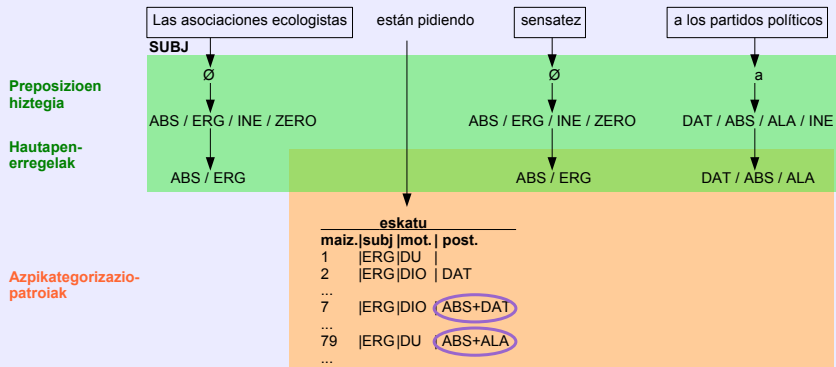
Hautapen-
erregelak



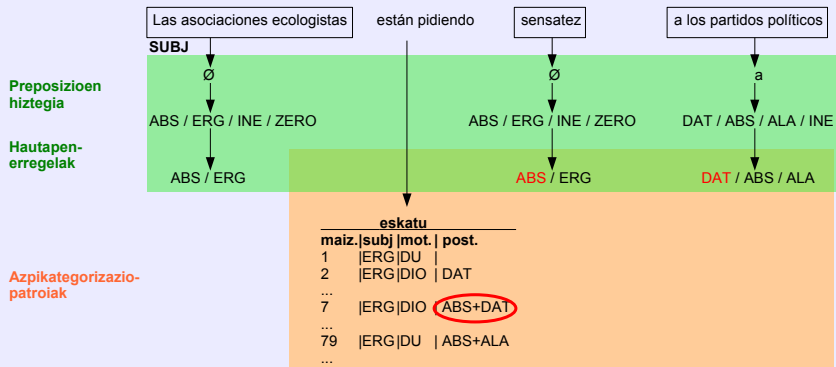
Preposizio eta funtzio sintaktikoen transferentzia. Adibidea



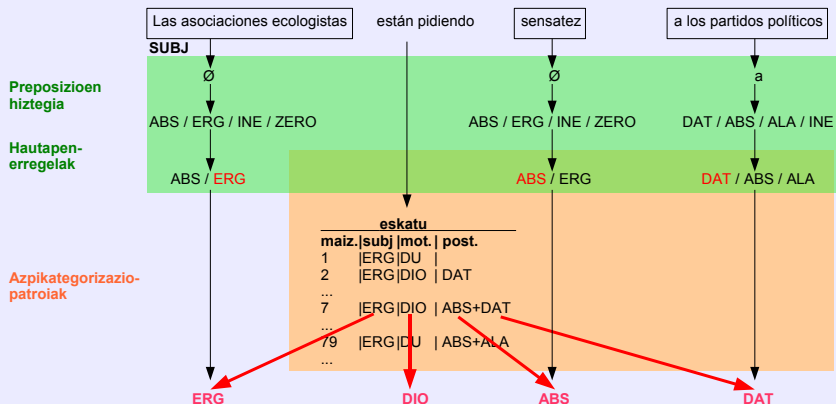
Preposizio eta funtzio sintaktikoen transferentzia. Adibidea



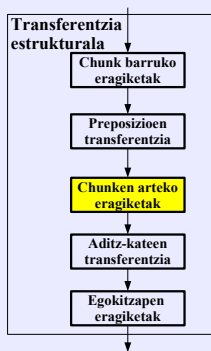
Preposizio eta funtzio sintaktikoen transferentzia. Adibidea



Preposizio eta funtzio sintaktikoen transferentzia. Adibidea



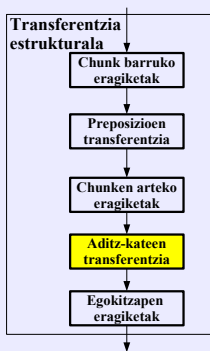
Chunken arteko eragiketak



Eragiketak:

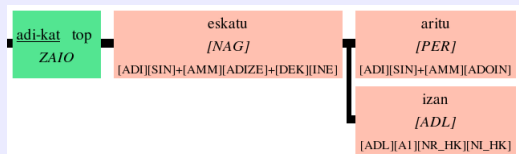
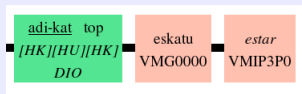
- Informazio-mugimenduak chunketik chunkera
 - Hurrengo modulueta beharko delako
- Nodorik ez duten chunken ezabaketa

Aditz-kateen transferentzia

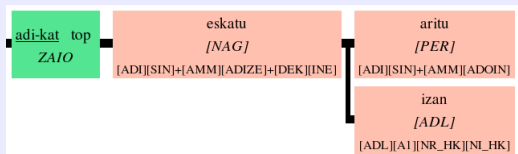
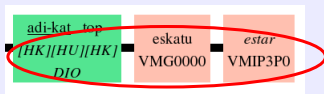


- Gramatika *XFST* sintaxi estandarrean oinarrituta
- Erregelak hiru multzotan antolatuta:
 - Aditz-kate mota identifikatzeko eta mota horri dagokion euskarazko eskema gehitzeko
 - Eskemako atributuak dagozkien balioekin ordezkatzeko
 - Soberazko informazioa garbitzeko

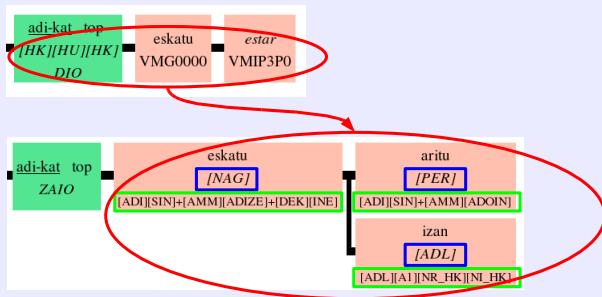
Aditz-kateen transferentzia. Adibidea



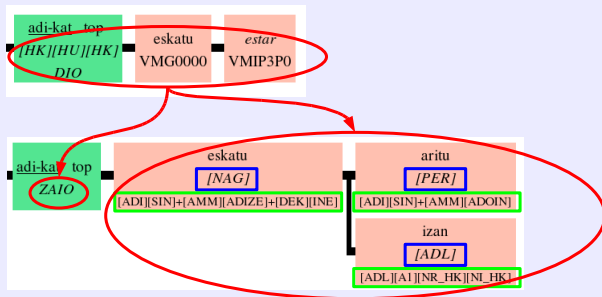
Aditz-kateen transferentzia. Adibidea



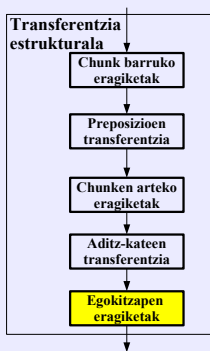
Aditz-kateen transferentzia. Adibidea



Aditz-kateen transferentzia. Adibidea



Egokitzapen-eragiketak

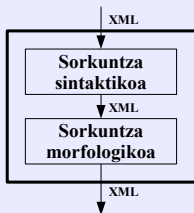


Bi moldaketa:

- Aditz-katearen mota aldatu bada subjektuaren postposizioa egokitu
Elkarte ekologist^{ak} zentzutasuna eskatzen ari zaizkie alderdi politikoei
- Aditz-chunkean mendekotasun-morfemarik badago aditz-laguntzaileari pasa

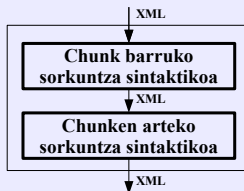
Sorkuntza

- Helburua:
 - Transferentzia-fasean lortutako xede-hizkuntzako egituratik testua ematea



Sorkuntza sintaktikoa

- Funtsa:
 - Hitzak eta osagaiak xede-hizkuntzako sekuentzia zuzen batean ordenatzen ditu
- Bi mailatan ordenatzen da:
 - Chunk barruko nodoak
 - Chunken artean

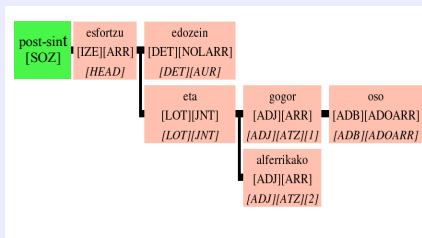


Chunk barruko sorkuntza sintaktikoa

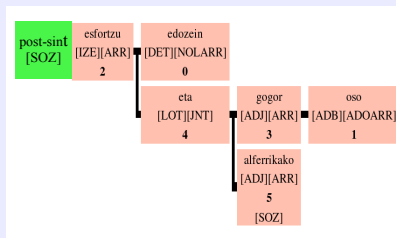
- Eginbeharrak:
 - Chunk barruan nodoen ordena erabaki
 - Euskaraz sintagma bakoitzaren elementuak modu jakin eta zurrun batean ordenatzen dira
 - Chunk mota bakoitzerako ordena hori aurrekotasun-erregelekin kodetu dugu
 - Postposizio-informazioa duten chunketan informazio hori chunkaren azkeneko nodoari pasa

Chunk barruko sorkuntza sintaktikoa. Adibidea

Con cualquier esfuerzo muy duro y baldío

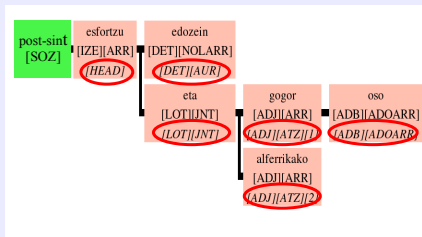


Edozein oso esfortzu gogor eta alferrikakoarekin

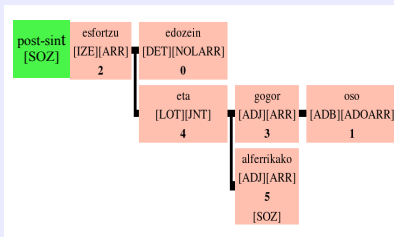


Chunk barruko sorkuntza sintaktikoa. Adibidea

Con cualquier esfuerzo muy duro y baldío



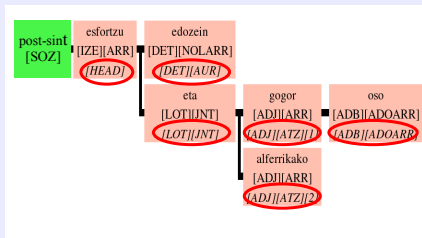
Edozein oso esfortzu gogor eta alferrikakoarekin



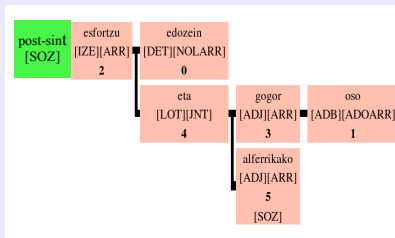
[DET] [AUR] → [ADB] [ADOARR] → [ADJ] [AUR] → [ADJ] [AUR] [1] → [LOT] [JNT] →
→ [ADJ] [AUR] [2] → [2] → [IZE] [IZB] → [IZE] [ARR] → [HEAD] →
→ [ADJ] [ATZ] → [ADJ] [ATZ] [1] → [LOT] [JNT] → [ADJ] [ATZ] [2] → [DET] [ATZ]

Chunk barruko sorkuntza sintaktikoa. Adibidea

Con cualquier esfuerzo muy duro y baldío



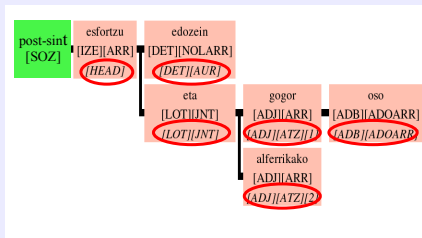
Edozein oso esfortzu gogor eta alferrikakoarekin



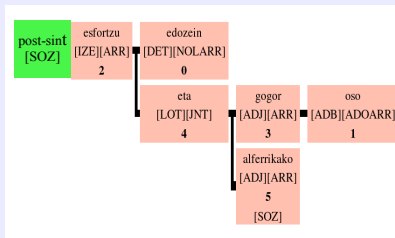
[DET] [AUR] → [ADB] [ADOARR] → [ADJ] [AUR] → [ADJ] [AUR] [1] → [LOT] [JNT] →
→ [ADJ] [AUR] [2] → [2] → [IZE] [IZB] → [IZE] [ARR] → [HEAD] →
→ [ADJ] [ATZ] → [ADJ] [ATZ] [1] → [LOT] [JNT] → [ADJ] [ATZ] [2] → [DET] [ATZ]

Chunk barruko sorkuntza sintaktikoa. Adibidea

Con cualquier esfuerzo muy duro y baldío



Edozein oso esfortzu gogor eta alferrikakoarekin



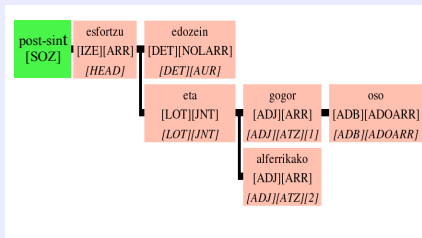
[DET] [AUR] → [ADB] [ADOARR] →

→ [HEAD] →

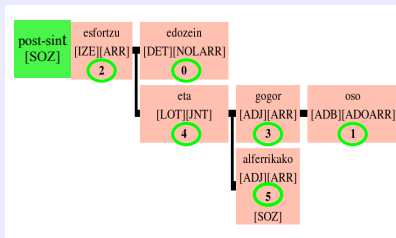
→ [ADJ] [ATZ] [1] → [LOT] [JNT] → [ADJ] [ATZ] [2]

Chunk barruko sorkuntza sintaktikoa. Adibidea

Con cualquier esfuerzo muy duro y baldío



Edozein oso esfortzu gogor eta alferrikakoarekin



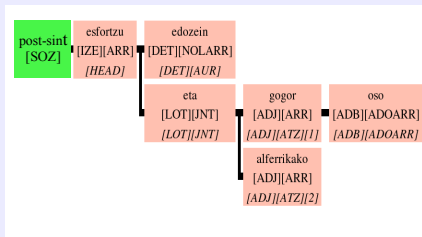
[DET] [AUR] → [ADB] [ADOARR] →

→ [HEAD] →

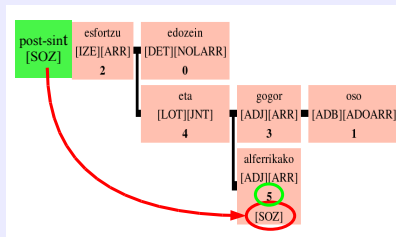
→ [ADJ] [ATZ] [1] → [LOT] [JNT] → [ADJ] [ATZ] [2]

Chunk barruko sorkuntza sintaktikoa. Adibidea

Con cualquier esfuerzo muy duro y baldío



Edozein oso esfortzu gogor eta alferrikakoarekin

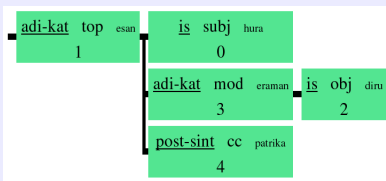
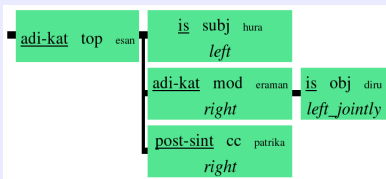
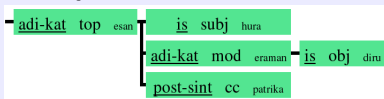


Chunken arteko sorkuntza sintaktikoa

- Bi urrats:
 - Ordena erlatiboa
 - Chunk bakoitza bere gurasoarekiko
 - Euskarazko perpauseko chunken arteko ordena oso librea bada ere, badira zenbait gomendio ordenatzeko (Zabala, 00). Aukera posibleen artean gure ustez orokorrean egokiena kodetu dugu
 - Erregeletan kontuan hartzen da: chunkaren mota, informazio sintaktikoa eta fokua, eta gurasoaren mota
 - Ordena absolutua
 - Ordena erlatiboen informazioa erabiltzen da

Chunken arteko sorkuntza sintaktikoa. Adibidea

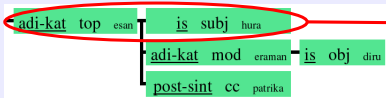
Ella dice que lleva dinero en el bolsillo



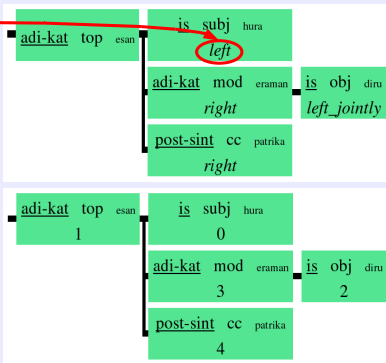
Hark esaten du dirua daramala patrikan

Chunken arteko sorkuntza sintaktikoa. Adibidea

Ella dice que lleva dinero en el bolsillo



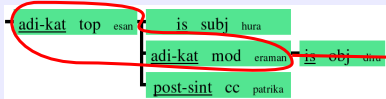
```
<rule id='C/C_relord_adi_sub'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat'] </def>
    <def> C2 := C1/CHUNK[@si='subj'] </def>
  </match>
  <actions>
    <act> C2/@relord := 'left' </act>
  </actions>
</rule>
```



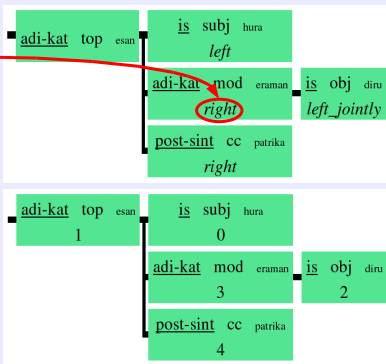
Hark esaten du dirua daramala patrikan

Chunken arteko sorkuntza sintaktikoa. Adibidea

Ella dice que lleva dinero en el bolsillo



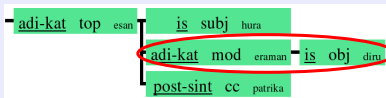
```
<rule id='C/C_relord_adi_bestela'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat'] </def>
    <def> C2 := C1/CHUNK[@focus!='true']
      [@si!='subj'][@type!='ez'] </def>
  </match>
  <actions>
    <act> C2/@relord := 'right' </act>
  </actions>
</rule>
```



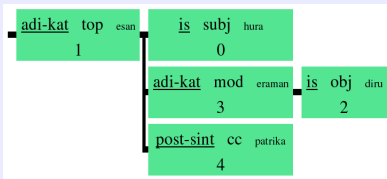
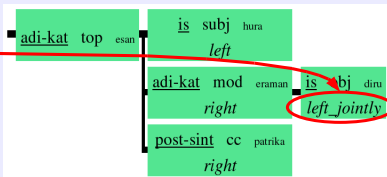
Hark esaten du dirua daramala patrikan

Chunken arteko sorkuntza sintaktikoa. Adibidea

Ella dice que lleva dinero en el bolsillo



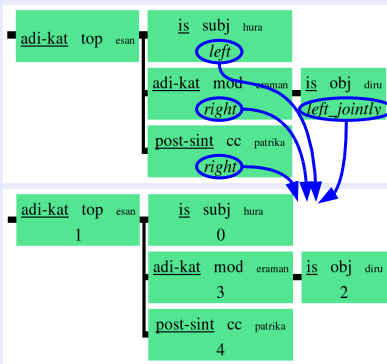
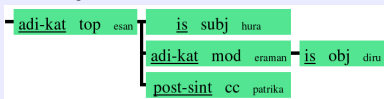
```
<rule id='C/C_relord_adi_focus'>
  <match>
    <def> C1 := //CHUNK[@type='adi-kat'] </def>
    <def> C2 := C1/CHUNK[@focus='true'] </def>
  </match>
  <actions>
    <act> C2/@relord := 'left-jointly' </act>
  </actions>
</rule>
```



Hark esaten du dirua daramala patrikan

Chunken arteko sorkuntza sintaktikoa. Adibidea

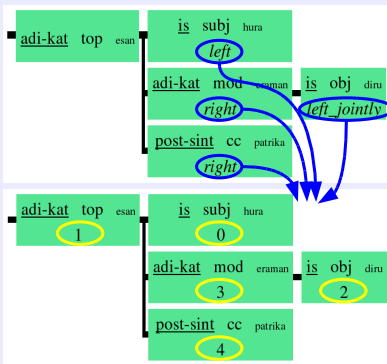
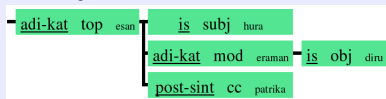
Ella dice que lleva dinero en el bolsillo



Hark esaten du dirua daramala patrikan

Chunken arteko sorkuntza sintaktikoa. Adibidea

Ella dice que lleva dinero en el bolsillo



Hark esaten du dirua daramala patrikan

Sorkuntza morfologikoa

- Helburua:
 - Xede-hizkuntzako hitzen formak sortzea
- *Morfeus* prozesadore morfologikoa (Alegria *et al*, 96)
berrerabili dugu:
 - Kode librekoa izateko egokitu
 - XMLn oinarritutako *Apertium* proiektuko hiztegien formatua
- Soilik postposizio-informazioa duten nodoak prozesatuko dira
 - Aditz-kateetan: nodo guztiak
 - Izen- eta preposizio-sintagmetan: azken nodoa
- Zifrak, datak eta orduak duten nodoetan sorkuntza morfologiko berezia
- Postposizio-informaziorik ez dagoenean: forma = lema

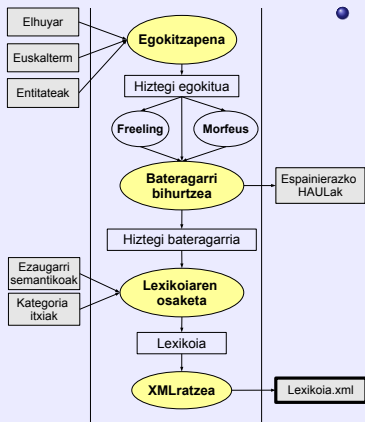
Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena
- 4 Baliabide linguistiko berrerabilpena**
- 5 Ebaluazioa
- 6 Ondorioak eta etorkizunerako lanak

Lexikoi elebidunaren eraikuntza

- Helburua: sarrera asko izango dituen lexikoi aberatsa eraikitzea, ahalik eta esfortzu txikienarekin → Berrerabilpena
- Baliabideak:
 - *Elhuyar* es-eu hiztegiaren bertsio elektronikoa:
 - Sarrerak (62.000)
 - Hitz anitzeko azpisarrerak (11.000)
 - *Euskalterm* terminologia banku publikoaren hitz anitzeko terminoak (1.700)
 - Corpusetatik erauzitako entitateen zerrenda elebiduna (910)

Lexikoi elebidunaren eraikuntza



Prozesaketa:

- 1 Egitura homogeneoa lortzeko egokitzapena
- 2 *Matxin*en beste baliabideekin bateragarri egitea
 - Katetoria-sistemak bateratu
- 3 Lexikoaren osaketa
 - Informazio semantikoa gehitu
 - Katetoria itxien hiztegia integratu
- 4 XML formatu estandarrera bihurtzea
 - *Apertium* proiektuaren espezifikazioari jarraituz

Ezaugarri semantikoaren etiketatze erdi-automatikoa

- Hitzen arteko azaleko erlazio semantikoak:
 - Genus, erlatorespezifikoak eta sinonimia
 - *Euskal Hiztegiaren* (Sarasola, 96) izenen definizioetatik ateratakoak (Agirre *et al.*, 00)
- Metodoa:
 - 1 Eskuzko etiketatzea: maiztasun handieneko genus eta erlatorespezifikoak
 - 2 Etiketatze automatikoa:
 - 1 Genus eta erlatorespezifikoak erabiliz
 - 2 Sinonimia erabiliz
 - 3 Etiketatze automatikoaren errepikapena

	[±bizidun] ezaugarria
Doitasuna	%99,2
Estaldura	%75,14

Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena
- 4 Baliabide linguistikoaren berrerabilpena
- 5 Ebaluazioa**
- 6 Ondorioak eta etorkizunerako lanak

Ebaluaziorako neurria

- Edizio-distantzia (Przybocki *et al.*, 06; Snover *et al.*, 06)
- Sistemaren itzulpenean giza-editore batek egin beharreko moldaketa kopurua:
 - abiapuntu-testuaren esanahi osoa edukitzeko
 - ulergarria eta gramatikalki zuzena izateko
 - ahalik eta moldaketa gutxien egiten
- *Moldaketa*:
 - Banakako hitzen:
 - txertatzea
 - ezabaketa
 - ordezipena
 - Hitz multzoen mugitzea

Edizio-distantzia

- Edizio-distantziak ebaluazio fidagarria ematen du:
 - Absolutua: irteera zenbaterainoko den erabilgarria neurtzeko
 - Erlatiboa: sistema desberdinen artean konparatzeko
- AEBetako DARPA agentziak erabilgarria dela frogatu du
- NIST erakundea erabiltzea aztertzen ari da, *BLEU* (Papineni et al., 02) bezalako neurrien eragozpenei aurre egiteko

Matxin 1.0 sistemaren ebaluazioa

- Ebaluaziorako corpusak:
 - *Eitb*: hizkuntza orokorreko kazetaritza-corpusa
 - *Consumer*: kontsumoaren arloko testuen corpusa
 - Corpus bakoitzetik 5 eta 25 bitarteko hitzetako 50 esaldi
- Edizio-distantziaren emaitzak:

	<i>Matxin 1.0</i>
Eitb	%40,4
Consumer	%43,6

- NIST erakundearen *MTEval04* kanpainen sistemen irteeraren edizio-distantzia (Przybocki *et al.*, 06; Snover *et al.*, 06)
 - %26 - %47

Bi sistemen konparazioa: *Matxin* vs *Matrex*

- *Matrex* (Stroppa & Way, 06)
 - Sistema hibridoa:
 - Adibideetan oinarritutako IA + teknika estatistikoak
 - Dublin City University-ko NCLT-MT taldeak garatuta
 - IXA taldearekin elkarlanean:
 - en→eu (Way *et al.*, 06)
 - es→eu (Labaka *et al.*, 07)
- *Matrex* es→eu
 - *Consumer* aldizkariko corpusarekin entrenatuta
 (es: 975.000 hitz; eu: 785.000 hitz)
- Ebaluazioaren emaitzak:

	<i>Edizio-distantzia</i>	
	<i>Matxin</i>	<i>Matrex</i>
Eitb	40,4	71,8
Consumer	43,6	57,9

Bi sistemen konparazioa: *Matxin* vs *Matrex*

- *Matrex* (Stroppa & Way, 06)
 - Sistema hibridoa:
 - Adibideetan oinarritutako IA + teknika estatistikoak
 - Dublin City University-ko NCLT-MT taldeak garatuta
 - IXA taldearekin elkarlanean:
 - en→eu (Way *et al.*, 06)
 - es→eu (Labaka *et al.*, 07)
- *Matrex* es→eu
 - *Consumer* aldizkariko corpusarekin entrenatuta
 (es: 975.000 hitz; eu: 785.000 hitz)
- Ebaluazioaren emaitzak:

	<i>Edizio-distantzia</i>		<i>Bleu</i>	
	<i>Matxin</i>	<i>Matrex</i>	<i>Matxin</i>	<i>Matrex</i>
Eitb	40,41	71,87	9,30	9,02
Consumer	43,60	57,97	6,31	8,03

Matxin sistemaren erroreen iturburu nagusiak

- Analisia

- Mendekotasun-analizatzailea
- Etiketatzaile morfosintaktikoa
- Funtzio sintaktikoen esleipena
- Hitz anitzeko terminoen identifikazioa

qué nota tiene? → *Du zerk **nabari du?***

- Hautapen lexikala

Pillan a un caco → *Xixkalari bat **lapurtzen dute***

- Preposizioen itzulpena

en el colegio de Rojas → *Rojales-**en** ikastetxean*

- ...

Itzulpen-adibideak

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
El acuerdo ha sido roto por los representantes	Akordioa ordezkariak hautsi dute
Los políticos piden que demos tiempo al tiempo	Politikariak eskatzen dute pazientzia izan dezagula

Itzulpen-adibideak

Los poderes públicos fomentarán la educación sanitaria, la educación física y el deporte	Botere publikoek osasun-hezkuntza sustatuko dute, gorputz-hezkuntza eta kirola
Fue entonces cuando escuchó la explosión que se produjo en el primer piso	Orduan izan zen leherketa entzun zue-nean 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
Como se sabe, muchos clientes habituales estudian minuciosamente las tarifas y optan por la que más conviene a su tipo de uso, cara a que su economía doméstica se vea lo menos afectada posible por este necesario gasto logístico	Jakin, ohizko bezero asko minuciosamente estudiantzen dituzte tarifak eta aukeratzen dute gehiago erabileraren haren tipoari bat datorkion, haren etxeko ekonomia gutxienez ikus dadila itxurati posiblea beharrezko gastu logistiko hau

Erabilera okerrak

The screenshot displays the website of the Ayuntamiento de San Sebastián. The page is in Basque. On the left, under the heading "En el Ayuntamiento", there is a paragraph about the city's commitment to improvement and a link to the "Equipo del Grupo Popular del Ayuntamiento de San Sebastián". On the right, a vertical navigation menu lists: "Inicio", "Contacta con M.J.U.", "En el Ayuntamiento", and "Acerca de M.J.U.". Below the main content area, there is a section titled "Udaletxean" (At the Town Hall) with a paragraph about the city's commitment to education and a link to "Donostiako Udalaren Alderdi Popularraren taldea". On the far right, another vertical navigation menu lists: "Hasiera", "MJUrekin kontaktatu.", "Udaletxean", "MJUren buruz", and "buruz". The Windows taskbar at the bottom shows the date as April 26, 2014, at 14:45.

En el Ayuntamiento

Nuestro firme compromiso por mejorar la ciudad, por consolidar su dinamismo y capacidad de liderazgo, por aumentar el grado bienestar de los donostiaras, es el principio rector de un sinfín de actuaciones de muy diversa naturaleza, encaminadas a construir entre todos el San Sebastián del futuro.

[Equipo del Grupo Popular del Ayuntamiento de San Sebastián](#)

Inicio
Contacta con M.J.U.
En el Ayuntamiento
Acerca de M.J.U.

Udaletxean

Gure konpromiso sendoa hobetu hiri, eta buruzagitzaren edukiera finkatu dinamismo haren, donostiarren handitu maila ongizate, asko natura desberdinaren jokaeren sinfin baten gidari hasiera da, eraikitze guztiak geroaren Donostia bideratu.

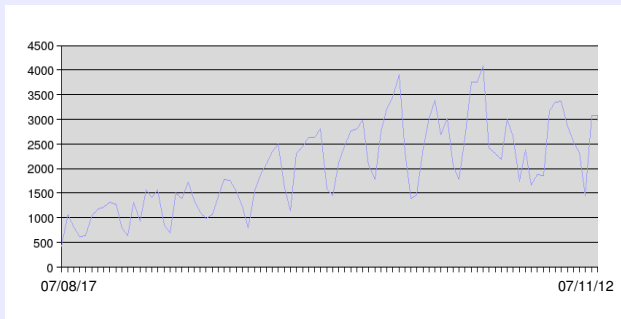
[Donostiako Udalaren Alderdi Popularraren taldea.](#)

Hasiera
MJUrekin kontaktatu.
Udaletxean
MJUren buruz

Erabilera okerrak



Itzulpen-eskaera



Aurkezpenaren eskema

- 1 Sarrera
- 2 Matxin itzulpen-sistemaren teknologia
- 3 Moduluen deskribapena
- 4 Baliabide linguistikoaren berrerabilpena
- 5 Ebaluazioa
- 6 Ondorioak eta etorkizunerako lanak

Ekarpenak

- *Matxin* euskarara itzultzeko transferentzia sakoneko sistemaren diseinua eta es→eu bikoterako inplementazioa
 - Mendekotasun-analizatzailea
 - Preposizio eta funtzio sintaktikoen transferentzia
 - Aditz-kateen transferentzia
 - Sorkuntza sintaktikorako moduluak
- Baliabide linguistikoaren berrerabilpena
 - Lexikoiaren eraikuntza estaldura handiko hiztegietatik
 - Ezaugarri semantikoaren etiketatze erdi-automatikoa
 - Prozesadore morfologikoaren egokitzapena
- Sistemaren ebaluazioa

Ekarpenak

- *Matxin* es→eu IAKo sistema:
 - Publikoki erabilgarria dago
<http://www.opentrad.org>
 - Kode irekiko software libre bezala banatzen da
<http://matxin.sourceforge.net>



Ondorioak

- Gure helburua bete dugu
 - Erregeletan oinarritutako itzulpen-sistema bat diseinatu eta inplementatu dugu
 - Sistemaren itzulpenek erregeletan oinarritutako estrategiaren
 - ahalmena frogatzen du
 - zailtasunak eta mugak identifikatzen ditu
- Gure hipotesia frogatu dugu
 - Posible da euskararako asimilaziorako IAKo sistema bat sortzea
 - Euskararako dauden baliabideekin, erregeletan oinarritutako estrategiak ahalmen handiak ditu

Ondorioak

- *Matxin* aitzindaria da:
 - Euskararekin lan egiten duen eta publikoki erabilgarria dagoen lehenengo IAKo sistema da
 - Software librean eraikitako lehenengo IAKo sistemetako bat da

Etorkizunerako lanak

- *Matxin* es→eu sistemaren hobekuntzak
 - Erroreen arazketa eta zuzenketa
 - Analizatzailearen hobekuntza
 - Diseinatutako erregelen idazketarako formalismo bateratuaren inplementazioa
- Teknika eta estrategia berrien ikerketa
 - Hautapen-lexikalerako desanbiguazio-teknikak
 - Domeinu zehatzetarako egokitzapena
 - Corpuetan oinarritutako teknikak
 - Hibridazioa
- Ataza berriak
 - Postediziorako interfazea
 - Hiztegia aberasteko tresnak
 - *Matxin* en→eu
 - Beste noranzko IAko sistemak: eu→es, eu→en

Matxin

Erregeletan oinarritutako itzulpen automatikoko
sistema baten eraikuntza
estaldura handiko baliabide linguistikoak berrerabiliz

Aingeru Mayor Martinez

Lengoaia eta Sistema Informatikoak Saila
Euskal Herriko Unibertsitatea

2007ko azaroaren 27a