2009ko apirilaren 29a

Udako Euskal Unibertsitatea

Morfologia eta sintaxiko ariketak konputagailuaren bidez

M. Arantzabe <maxux.aranzabe@ehu.es>
K. Sarasola <kepa.sarasola@ehu.es>

Aurkibidea

1 Sarrera

1.1 Motibazioa

Linguistika konputazionalaren arloak informazio linguistikoarekin egiten du lan. Izan ere, konputagailuak linguistikari eskaini dion ekarpen nagusia testu handiekin hainbat datu ateratzea baita, datu horiek analizatu, eta analisi horretatik hainbat ondorio ateratzeko aukera emanez. Egun hainbat aukera dira analisi morfologiko eta sintaktikoa automatikoki lortzeko. Tresna horietako batzuk Interneten erabil daitezke publikoki hitzen analisi morfologiko eta esaldien analisi sintaktikoa automatikoki egiteko, Morfeus, Ancora eta Freeling adibidez. Tresna horiek eskolako edota institutuko ikasleentzat sintaxia eta morfologia ordenagailuarekin lantzeko erakargarriak izan daitezke, zuhaitzak modu grafikoan ikusten direlako edo hainbat proba di-da egin daitezkeelako.

1.2 Helburuak

Oinarrizko ariketa bilduma sortu ahal izateko zenbait baliabide aurkeztu nahi ditugu ikastaro honetan. Egun hainbat aukera dira Interneten analisi morfologiko eta sintaktikoa automatikoki lortzeko. Ezagutu behar dira guneak eta emaitzak erakusteko moduak, alegia erabiltzen diren formatuak, egiturak, kategoriak eta azpikategoriak. Hauek dira ikastaroaren helburu zehatzak:

- Euskarazko hitzak eta esaldiak Internet bidez, analizatzen jakitea.
- Analisien emaitzetan azaltzen diren kategoriak ondo ezagutzea.
- Hainbat ariketa egitea eta ariketa berriak asmatzeko erraztasuna lortzea.
- Gaztelaniaz eta ingelesezko ariketekin ere antzera.

1.3 Metodologia

Aztertuko dira Interneten bidez analisi morfologiko eta sintaktikoa egiteko dauden zenbait aukera. Emaitzak interpretatzeko behar diren kategoria eta azpikategoriak ondo ezagutzeko hainbat ariketa egingo dira konputagailuan. Bukaeran ikastaroko partaideek eurek asmatu beharko dituzte ariketa berriak, gero beren ikasleekin gelan erabili ahal izango dituztenak.

Batez ere euskararako egingo da, baina espainiera eta ingelesa ere landuko dira.

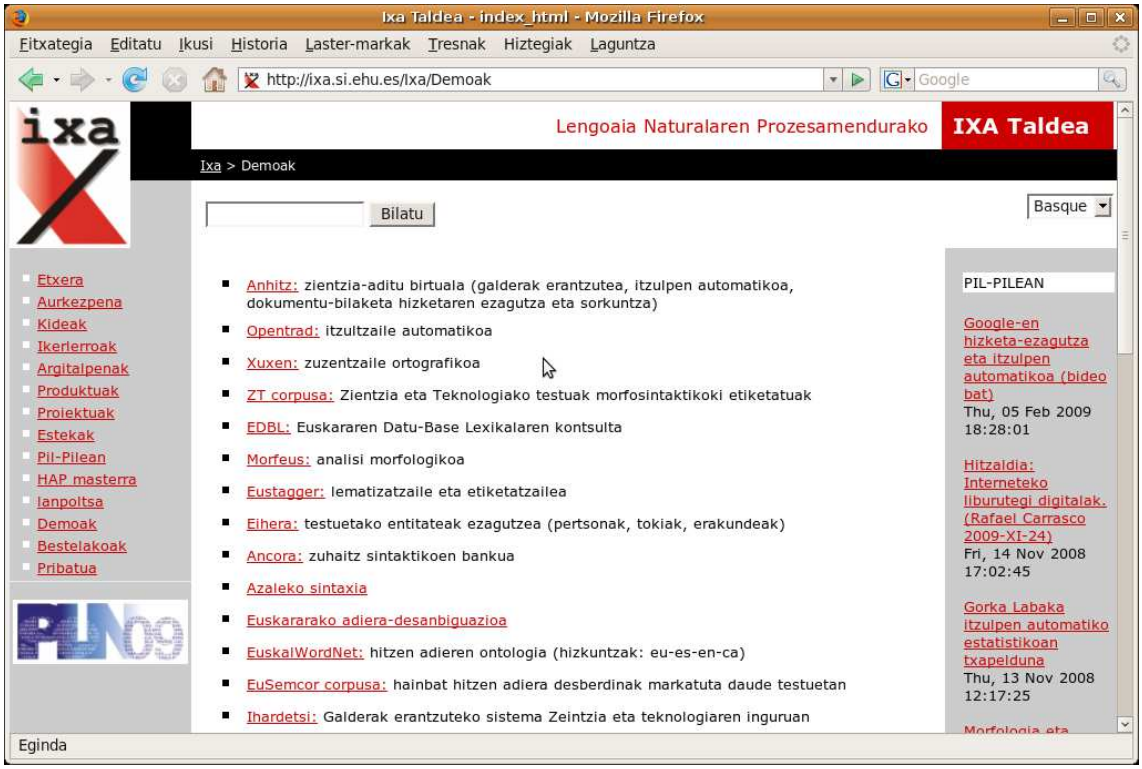
2 Euskararen morfologia lantzen: Morfeus, hitza isolatuta.

2.1 Analisi morfologikoa martxan: Morfeus

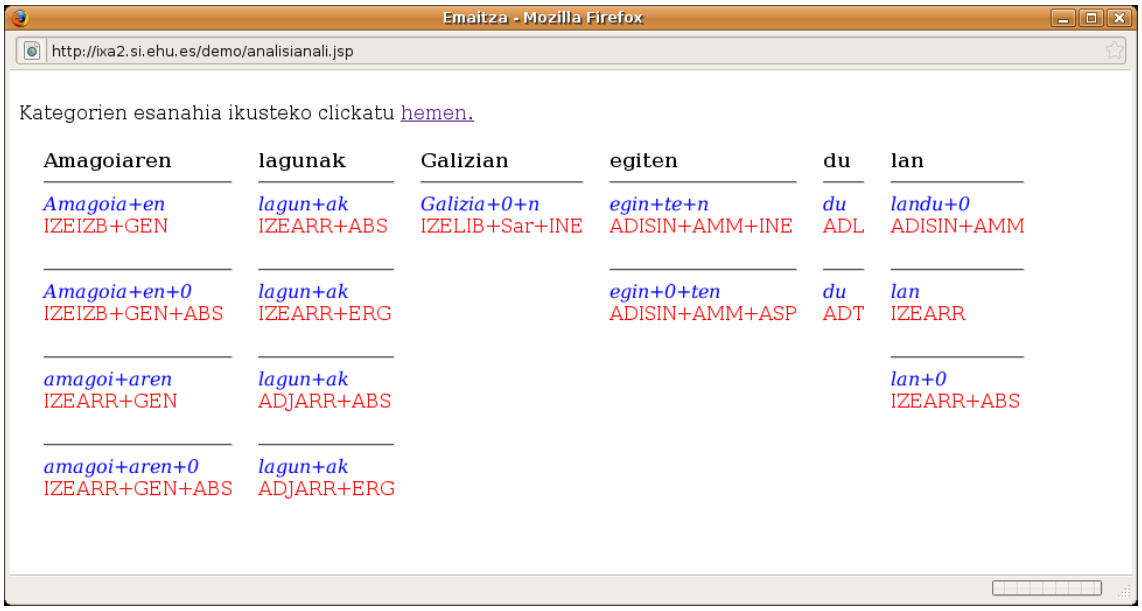
Bilatu Morfeus, euskarako analizatzaile morfologikoaren demoa Interneten. Horretarako jo ezazu zuzenean helbide honetara:

http://ixa2.si.ehu.es/demo/analisianali.jsp

Edo IXA taldearen orritik (ixa.si.ehu.es) abiatuta, aukeratu *Demoak* ezkerreko menuan, bertan aukeratu "*Morfeus: analisi morfologikoa*" eta lehen eman dugun helbide berera ailegatuko zara.



Morfeus analizatzaile morfologikoa zuzenean nola dabilen ikusiko dugu. Orri horretan zaudela idatz ezazu 'Analisi Morfologikoa' leihotxoan "*Amagoiaren lagunak Galizian egiten du lan.*" esaldia eta sakatu *Analizatu* botoia. Honako leiho bat zabalduko da hitz bakoitzak izan ditzakeen analisi morfologiko guztiekin:



Analisi bakoitzean bi lerro azaltzen dira, batean letra urdinetan ikusten dira hitzaren barruan identifikatu diren lema eta flexio-atzizkiak (adibidez: Amagoia+en), eta bigarren lerroan, letra gorritan, hitzaren informazio morfologikoa ikusten da hurrenez hurren (adibidez: IZEIZB+GEN; hau da, *Amagoia* IZEn kategoriako hitza dela, zehatzago esanda, izen berezia (IZB) eta *en* atzizkia, GENitibo kasuaren erlazio-atzizkia daramala).

Ondoko atalean erakusten da zein diren kategoria nagusiak. Kategoria nagusiaz gain azpikategoria ere erakusten da analisian, esate baterako IZEARR eta IZEIZB azaltzen dira analisietan. Kasu bietan, kategoria nagusia izena da (IZE), baina kasu batean azpikategoria ARR da (IZEARR: izen arrunta) eta bestean IZB (IZEIZB: izen berezia).

2.2 Kategoria multzoa

Euskarazko hitzak morfologikoki bereizteko Morfeusek erabiltzen dituen kategoria nagusiak hauek dira:

- IZE izenak
- ADJ adjektiboak
- ADI aditzak
- ADB adberbioak
- DET determinatzaileak

- IOR izenordainak
- LOT loturazkoak
- PRT partikulak (omen, ote...)
- ITJ interjekzioak (alajaina!)
- BST bestelakoak (baldin)
- ADL aditz laguntzaileak (du)
- ADT aditz sintetikoak edo trinkoak (dator)
- SIG siglak (EHU)
- SNB sinboloak (km, cm, g...)
- LAB laburdurak (etab.)

Kategoria-sistema osoa lehenengo eranskinean dago azalduta, 9.1. atalean, alegia; hor ikus daitezke kategoria eta azpikategoriak. Ixa taldeko demoetan ere ikus daitezke kategoria horiek, emaitzak erakutsi eta gero esteka bat eskaintzen baita kategoria guztiak beste leiho berri batean erakusteko (http://ixa2.si.ehu.es/edblkontsulta/labur-eus.htm).

2.3 Ariketak

A. Analizatu morfologikoki Morfeus erabiliz honako hitzak eta esaldiak:

1. Itxura hori zuen gizonak ikusi du.
2. Haurrak bizkor esan zuen etorriko zela.
3. duen, zioen, nuen, gazte, gorri, gaur, ikusten, ikus, alajaina, omen.
4. Nik huts egiten ez baldin badut, aurten ekarriko ditugu.
5. Bart Paris hoteleko zure gela arakatu egin dute
6. Gaur izebari omenaldia egin diote herrian.
7. *Nik iaz Parisera joan nintzen.
8. Nik iaz Parisera joateko aukera aprobetxatu egin nuen.

B. Xuxen zuzentzaile ortografikoak “or” hitza ez du gaizkitzat hartzen, zergatik? Ez luke gorriz azpimarratu behar? Letra bat behar ote luke hasieran? Aztertu ea analisi posiblerik dagoen “or” hitzerako. Berdin joka zenezake Xuxenek ontzat hartzen dituenekin nahiz eta zuretzat txarto egon.

3 Euskararen morfologia lantzen: Eustagger, hitza bere testuinguruan

3.1 Eustagger martxan

Aurreko atalean *"Amagoiaren lagunak Galizian egiten du lan."* esaldia analizatzean, hitz bakoitzerako analisi posibleak lortu ditugu. Analisi horietatik guztietatik zein da egokiena, ordea?

Hori da hain zuzen ere, Eustagger lematizatzailearen bitartez lortuko duguna. Horretarako, bilatu Eustagger lematizatzailearen demoa Interneten. Jo ezazu zuzenean helbide honetara: http://ixa2.si.ehu.es/demo/analisimorf.jsp. Edo lehen bezala, IXA taldearen orritik (ixa.si.ehu.es) abiatuta, aukeratu *Demoak* ezkerreko menuan eta bertan aukeratu "*Eustagger: lematizatzailea eta etiketatzailea*". Lehen eman dugun helbide berera ailegatuko zara: http://ixa2.si.ehu.es/demo/analisimorf.jsp.

Bertan ageri den leihotxoan lehengo esaldi bera jartzen badugu ("Amagoiaren lagunak Galizian egiten du lan."), orain lematizatzailearekin lortuko dugun emaitza ez da analizatzailearekin lortzen genuen bera. Hitz bakoitzak analisi bakarra emango du, ondoko irudian ikus daitekeen bezalaxe.

Gizakiontzat oso erraza da geure hizkuntza ulertzea, konputagailuari asko kostatzen zaio, ordea. Adibidez, testu bateko hitzak irakurtzen ditugunean, guk ez ditugu kontuan hartzen ezohiko diren interpretazio bitxiak, baina konputagailuak bai, denak aztertu behar ditu eta. Programa lematizatzaileek laguntzen diote konputagailuari perpaus bateko hitz bakoitzari dagokion interpretazio morfologiko egokia aukeratzen.

Argi dago Morfeus analizatzaileak hitz bakoitza testuingurua kontuan hartu gabe analizatzen duela. Erabili dugun perpausean *lagunak* hitza adjektiboa ere izan daitekeela

dio; *du* hitza aditz trinkoa ere izan daitekeela, nahiz eta aurreko hitza *egiten* izan; edo *ikusi* hitza izena.

Geroago analizatu dugu esaldi bera lematizatzailearekin. Lematizatzaileak analisi morfologikoa egiten du, baina hitzaren testuingurua aztertuta hitz bakoitzerako analisi bakarra aukeratzen du.

Morfeus analizatzaile morfologikoak batez beste euskarazko hitz bakoitzerako 2,81 analisi diferente ematen ditu. Kategoria eta azpikategoria sintaktikoa bakarrik kontuan hartuta, kasuak eta beste ezaugarri morfologikoak kontuan hartu gabe, 1,5 analisi ematen du hitz bakoitzeko, batez beste. Lematizatzaileak, ordea, testuingurua aztertu ondoren, lema eta kategoria bakarra hautatzen du hitz bakoitzerako. Hanka sartzen du, baina % 1 edo % 2an baino ez. Oso tresna erabilgarria da hizkuntza-teknologian.

3.2 Ariketak

A. Analizatu Eustagger lematizatzailea erabiliz honako esaldiak:

1. Itxura hori zuen gizonak ikusi du.
2. Haurrak bizkor esan zuen etorriko zela.
3. duen, zioen, nuen, gazte, gorri, gaur, ikusten, ikus, alajaina, omen.
4. Nik huts egiten ez baldin badut, aurten ekarriko ditugu.
5. Lodia ez izateak itxuraren lerdentasuna azpimarratzen zion.
6. Nik huts egiten ez baldin badut aurten ekarriko ditugu.
7. Bart Paris hoteleko zure gela arakatu egin dute.
8. Gaur izebari omenaldia egin diote herrian.
9. *Nik iaz Parisera joan nintzen.
10. Nik iaz Parisera joateko aukera aprobetxatu egin nuen.

4 Euskararen sintaxia lantzen: AnCora, Zatiak eta Eihera

Analisi sintaktiko automatikoa ez dabil analisi morfologiko eta lematizazioa bezain zorrotz. Zaila da automatikoki lortzea esaldi baten benetako analisi sintaktikoa. Analisi sintaktikoa egiteko hainbat aukera ikusten ditu konputagailuak, baina esaldia bera ulertu gabe zaila da jakitea zein den analisi egokia. Horregatik, analisi sintaktikoa aztertzeko atal honetan, bi bide hartuko ditugu:

- *AnCora*. Euskarazko, gaztelaniako eta katalanezko corpusa da, hainbat hizkuntza mailatan etiketatua. Euskarazkoari dagokionez, bertan 3.175 esaldiren analisia kontsulta daiteke. Esaldi horien analisi sintaktikoak automatikoki egin ziren, baina gero IXA taldeko hizkuntzalariek eskuz desanbiguatu zituzten, esaldi bakoitzaren analisi egokia aukeratuz. *AnCora*n ezingo duzu edozein esaldi analizatu, zuk asmatutako edozein esaldi, alegia. Eskuz landu diren 3.175 esaldi horiek bakarrik ikusi ahal izango dituzu. Hori bai, zuhaitzak grafikoki oso erakargarria den modu batean kontsultatu ahal izango dituzu.
- *Zatiak* eta *Eihera* sintaxia lantzeko bi tresna dira. *Zatiak* esaldi bateko sintagma modukoak bereizten ditu (*chunk* edo *zati* ere esaten zaie sintagma osoak baino txikiagoak izan daitezkeen hitz-kate horiei). *Eiherak*, berriz, esaldi batean agertzen diren entitateak identifikatzen ditu, baita beraien artean bereiztu ere tokia, pertsona edota erakundea adierazten dituztenak. *AnCora*rekin konparatuz, *Zatiak* eta *Eihera* tresnekin edozein esaldi azter daiteke, baita guri bururatzen zaigun edozein ere; baina noski soluzio bakarra ematen duenez, soluzio hori beti ez da egokia, tresnak automatikoki aukeratu behar izan du analisi bat nahiz eta esaldiaren esangura ez ulertu, beraz, ulergarria da beti ez asmatzea.

4.1 Ancora: esaldi analizatuak kontsultatzen

Esan bezala, AnCora euskarazko, gaztelaniazko eta katalanezko corpusa da, hainbat hizkuntza-mailatan etiketatua:

* Kategoria morfologikoa

* Osagai eta funtzio sintaktikoak

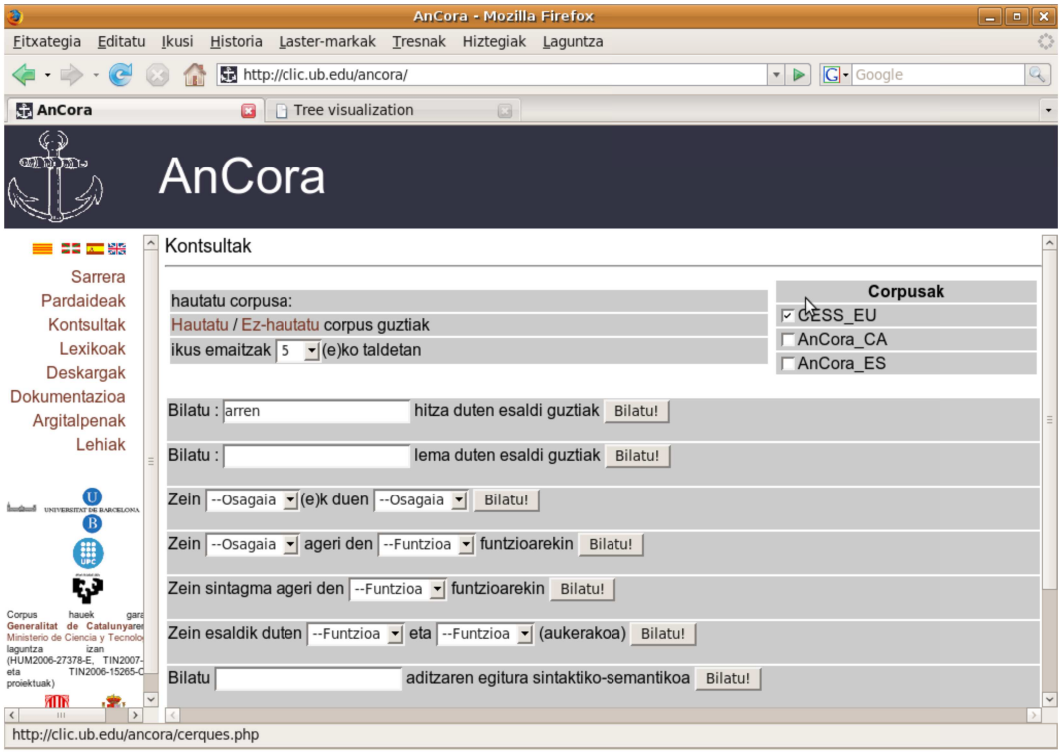
* Wordneteko adierak izenetan

* Entitateak

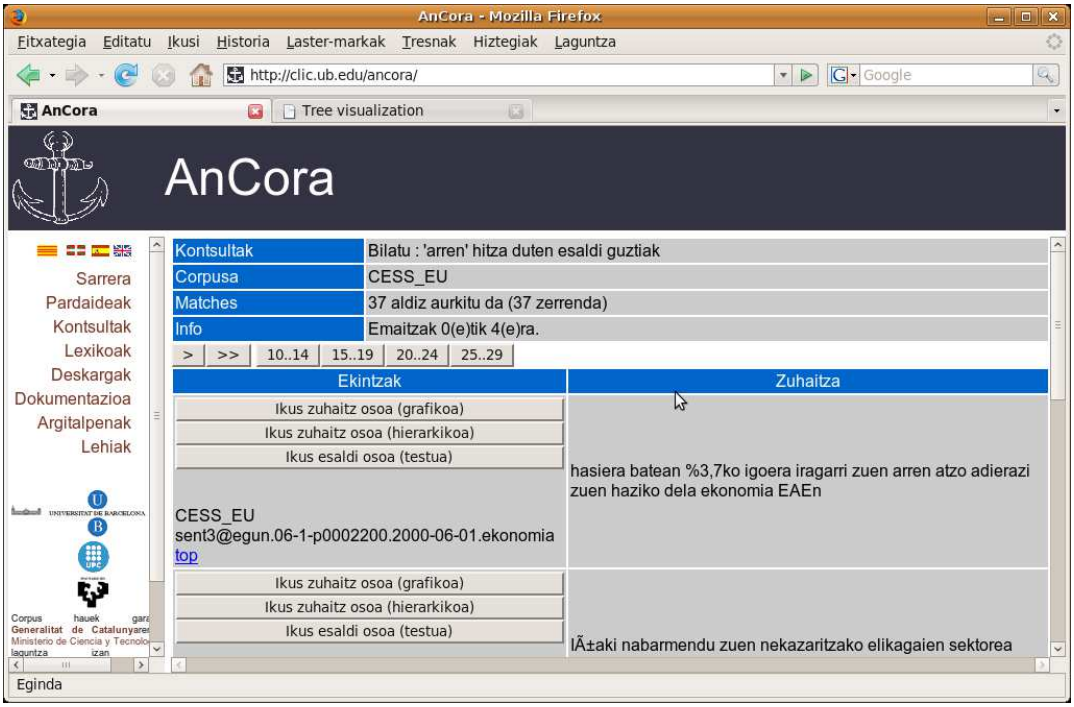
Gaztelaniaz eta katalanez 500.000 hitz inguru eskaintzen dira AnCora corpusean, eta euskaraz 55.000 hitz inguru. Euskarazko AnCora corpus etiketatuaren oinarria, EPEC corpusa (Aduriz eta al., 2006) da. EPEC, euskararen tratamendu automatikorako erreferentzia-corpus gisa darabilgu Ixa taldean. Heren bat XX. mendeko euskararen corpus estatistikoari dagokio eta beste bi herenak Euskaldunon Egunkariari. AnCora-rako 55.000 hitz hautatu dira, CESS-ECE proiektuan garatu zen sintaktikoki etiketatutako corpusaren zati bat, eta dependentzia-eredutik osagai-eredura pasatu dira. Corpus-zati honen % 25 katalanezko eta gaztelaniazko corpusekin konparagarria da, aldi bereko berriak jasotzen baitira.

Corpus hau nola balia daitekeen ikusteko, bilatu euskarazko, katalanezko eta espainierazko testu-corpus analizatuak kontsultatzeko AnCora webgunea. Jo ezazu zuzenean helbide honetara: http://clic.ub.edu/ancora/index.php. Edo lehen bezala, IXA taldearen orritik (ixa.si.ehu.es) abiatuta, aukera ezazu *Demoak* ezkerreko menuan eta bertan aukeratu "Ancora: *zuhaitz sintaktikoen bankua*".

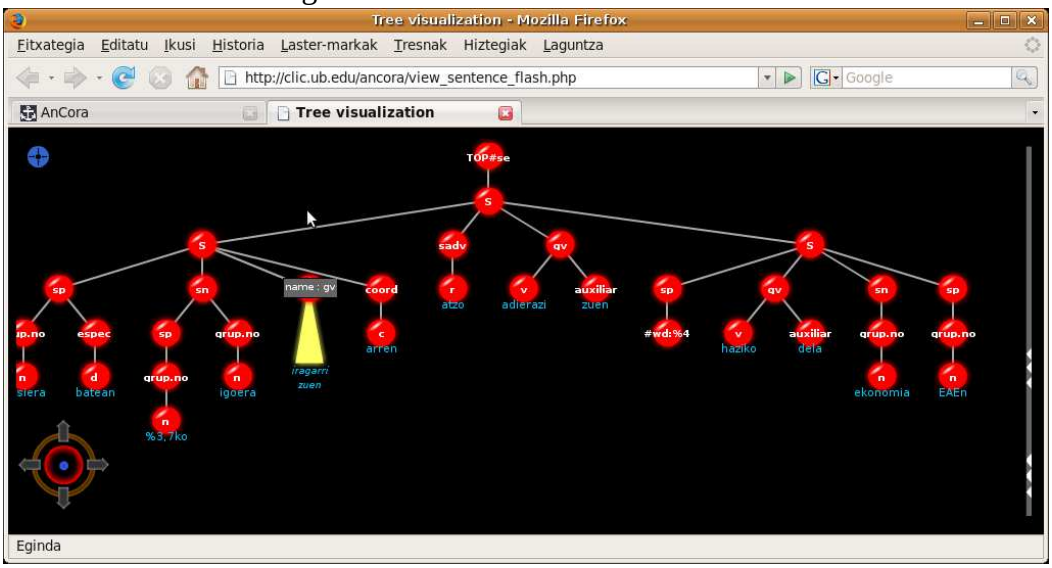
Ancora zerbitzuaren hasierako web-orrian aurkezpena eta hainbat datu ematen dira. Ezkerreko menuko Kontsulta aukeratuz honako pantaila batera pasatuko gara:



Hor, eskuinaldean, goian CESS_EU corpusa aukeratuko dugu eta gero nahi dugun esaldia bilatu beharko dugu. Pantaila horretako adibidean, "*arren*" hitza duten esaldiak bilatzeko eskatu da. Ondoko pantaila azalduko da gero bilaketaren emaitzak aurkezteko:

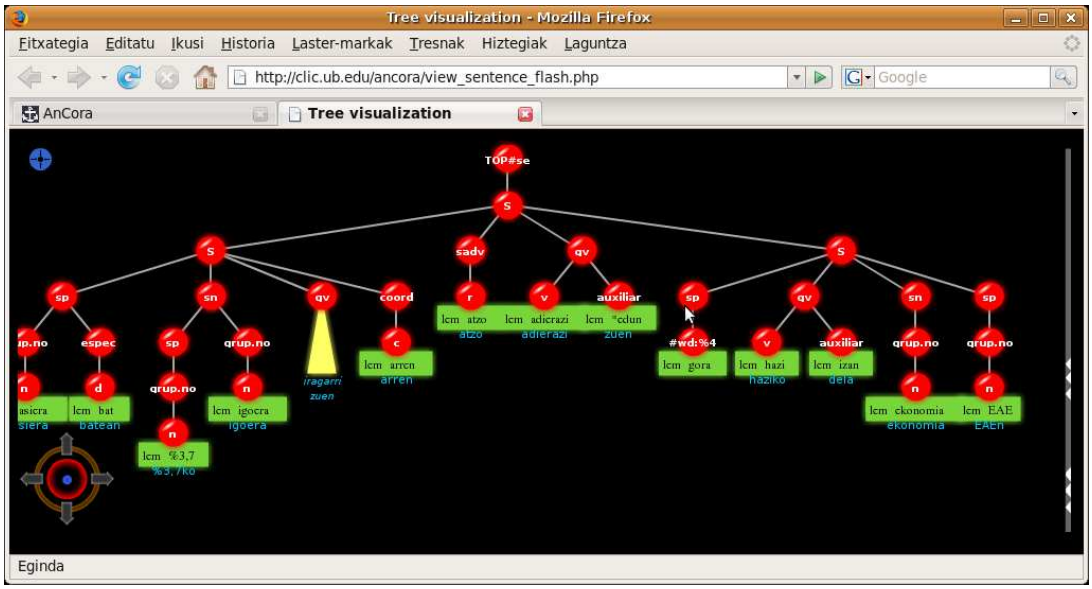


Demagun aztertu nahi dugula lehenengo esaldiaren analisia ("*Hasiera batean %3,7ko igoera iragarri zuen arren atzo adierazi zuen haziko dela ekonomia EAEn*") bere ezkerrean dagoen botoien artean "*ikus zuhaitz osoa (grafikoa)*" aukeratuz gero ondoko analisia ikusiko dugu:



Orain hainbat aukera ditugu zuhaitz horretan dagoen informazioa hobeto ikusteko. Aukera praktikoenak hauek dira:

- Nodo batean klikatuz gero bere azpian dagoen adar osoa trinkotu egingo da. Aurreko irudiko zuhaitzean horrela egin da "*iragarri zuen*" hitzen adarraren irudia trinkotuz.
- Nodo bakoitzean zer ezaugarri erakutsi behar den zehatz dezakegu eskuinean azaltzen diren menuak erabiliz. Ondoko irudian ikus daiteke nola erakusten den zuhaitza nodo guztien lema aurkezteko eskatuz gero.



4.2 Ariketa:

A. Bilatu itzazu Ancoran esaldi hauen analisia, eta konpara itzazu emaitzak Morfeus eta Eustagger-ek ematen dituzten analisiekin.

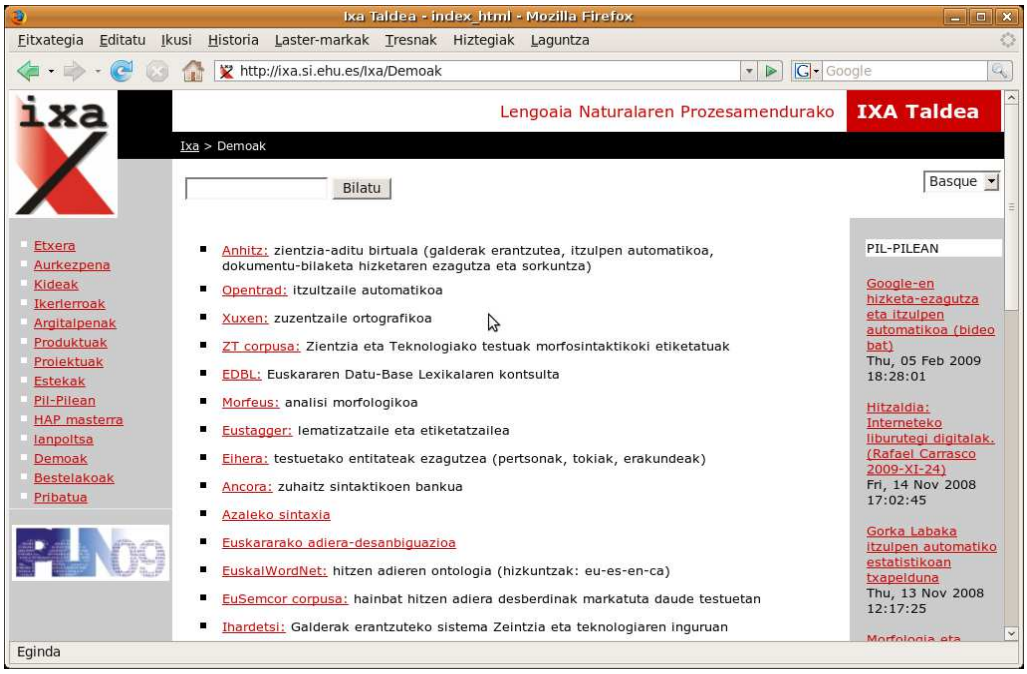
1.Lodia ez izateak itxuraren lerdentasuna azpimarratzen zion.

2.Nik hutsegiten ez baldin badut aurten ekarriko ditugu.

3.Bart Paris hoteleko zure gela arakatu egin dute.

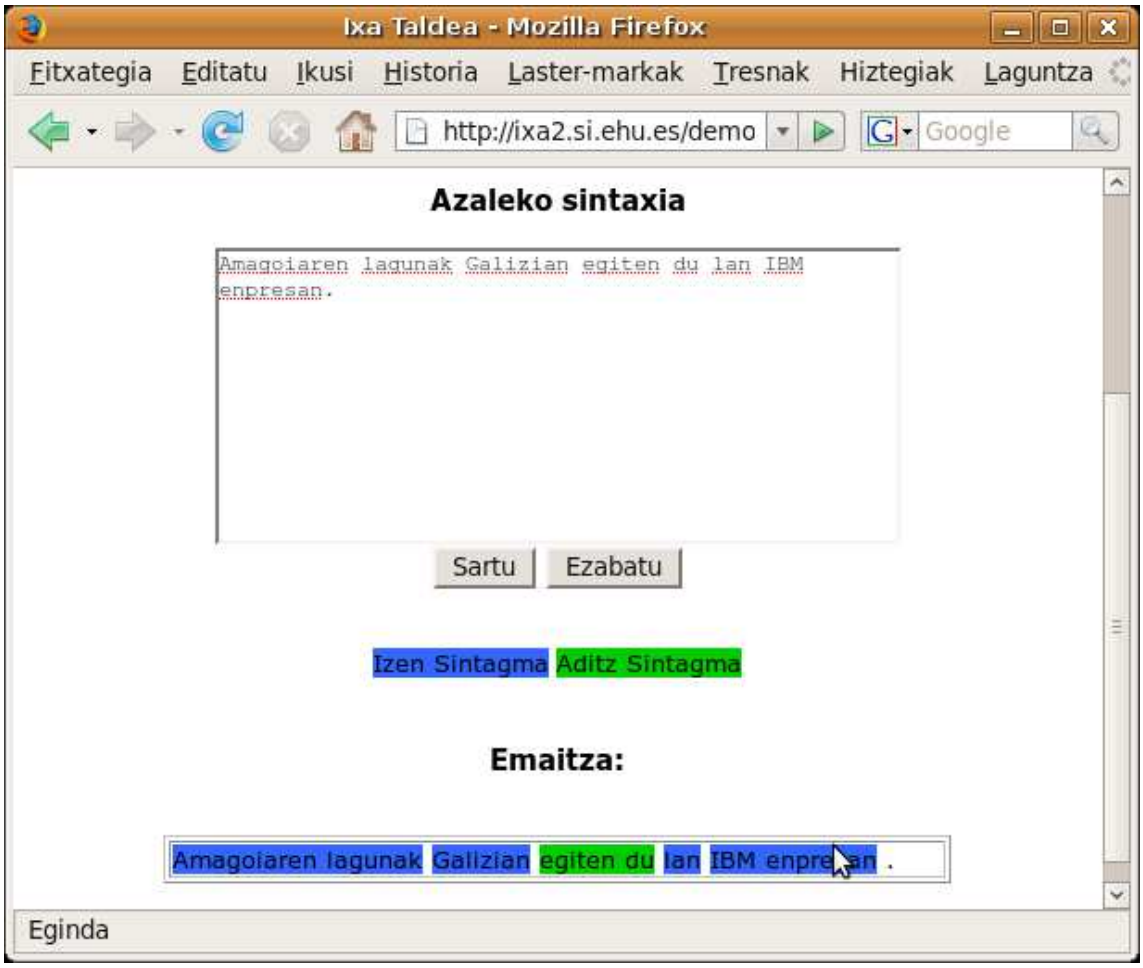
4.3 *Zatiak* (azaleko sintaxia) eta *Eihera* (entitateak) sintaxi-tresnak

Zatiak eta Eihera sintaxia lantzeko bi tresna dira. Tresna horiek erabiltzeko nahikoa da IXA taldeko demoen web-orrira joatea eta Azaleko sintaxia edo Eihera: esteketan klikatzea.



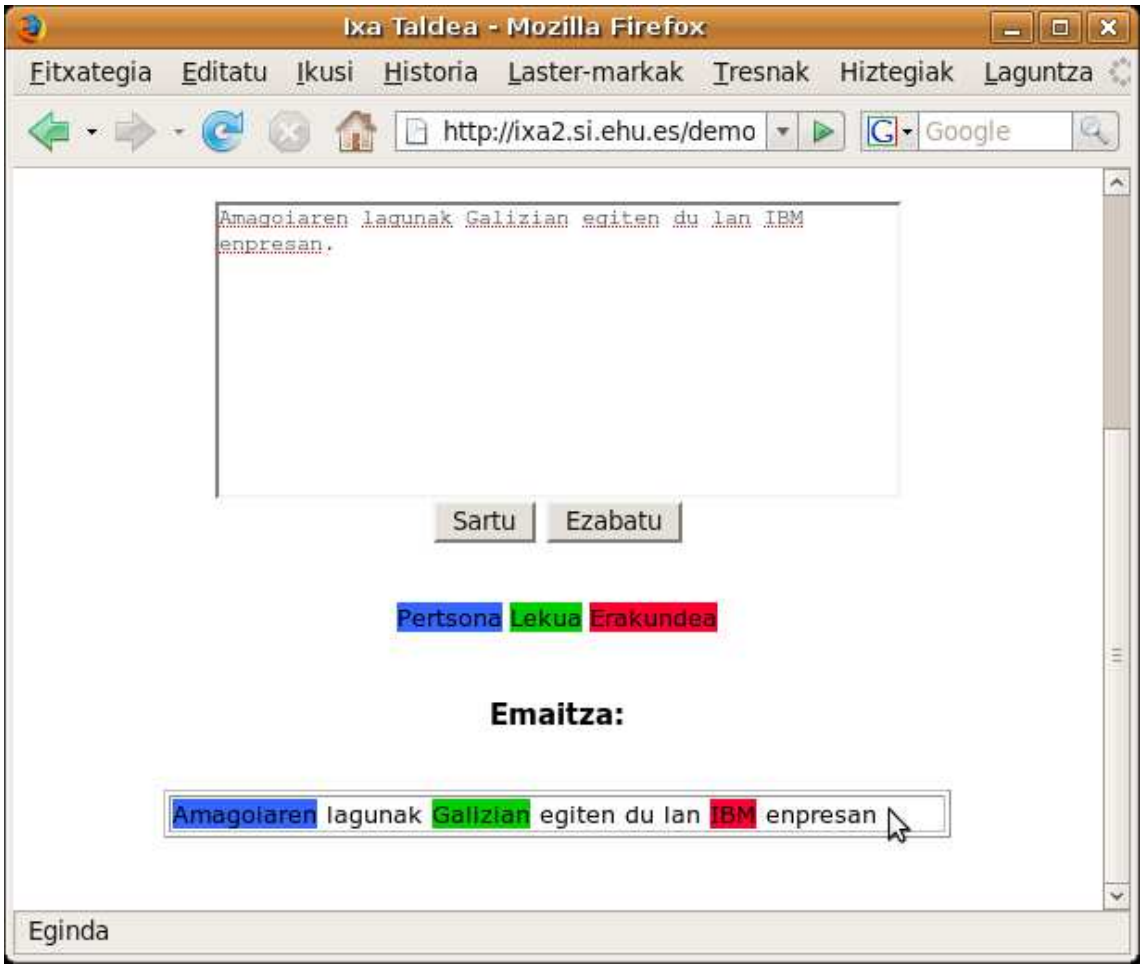
Zatiak programak esaldi bateko sintagma modukoak bereizten ditu. *Chunk* edo *zati* ere esaten zaie sintagma osoak baino txikiagoak izan daitezkeen hitz-kate horiei. Analisi sintaktiko osoa ez da lortzen horrela, zati horien arteko harremana ez baita zehazten. Azaleko sintaxia edo sintaxi partzial (shallow parsing) deritzogu analisi mota honi.

Azaleko sintaxia demoaren orrian zaudela, idatz ezazu *"Amagoiaren lagunak Galizian egiten du lan IBM enpresan"* eta sakatu *Sartu* botoia. Honako leiho bat zabalduko da esaldi bat osatzen duten sintagma edo kate sintaktiko guztiekin:



Azterturiko esaldi horretan, urdinez adierazi dira esaldiko sintagmak (<*Amagoaiaren lagunak*>, <*Galizian*>, <*lan*> eta <*IBM enpresan*>) eta berdez aditz sintagma (<*egiten du*>). Zehatz-mehatz hitz eginda horiek denak ez dira izen-sintagmak, <*Galizian*> eta <*IBM enpresan*> preposizio-sintagma direla esan beharko litzateke, baina zentzu orokor batean hartuta horiek ere izen sintagma gisa hartzen ditu *Zatiak* programak.

Eihera tresnak, berriz, esaldi batean agertzen diren entitateak (pertsonak, tokiak, erakundeak) identifikatzen ditu. Ikus dezagun, bada, nola egiten duen lan. Behin *Eiheran* zaudela, idatz ezazu aurreko esaldi bera, *"Amagoiaren lagunak Galizian egiten du lan IBM enpresan"*, eta sakatu *Sartu* botoia. Honako leiho bat zabalduko da zeinetan entitateak kolore desberdinez markatuak agertuko diren: urdinez, pertsona-izenak (<*Amagoaiaren*>); berdez, leku-izenak(<*Galizian*>); eta gorriz, erakunde-izenak (<*IBM*>).



4.4 Ariketa

A. Bereizi sintagmak ondoko esaldietan Zatiak tresna erabiliz:

1. Hondarribiako arrantzaleak kofradian bildu ziren atzo.
2. Peru baserriko amonarengana joan da.
3. Gazte horien esamesak ez dira niretzat oso fidagarriak.
4. Gezur hori esan zenien lagunei hementxe bertan gure aurrean.
5. Haurraren jostailu berri hauek egurrezkoak dira.

6. Ikastolan, ikasleek astero hainbat gai desberdini buruzko azalpenak egiten dituzte.
7. Su handiak urrutiko mundu magiko batera eramango zaitu.
8. Azterketari zaila izango zelako susmoa hartu genion.

B. Identifikatu entitateak ondoko esaldietan Eihera tresna erabiliz:

1. Hondarribiako arrantzaleak kofradian bildu ziren atzo.
2. Peru baserriko amonarengana joan da.
3. 1952.eko maiatzean Luis Villasante Kortabitarte jaunak bere sarrera-hitzaldia egin zuen Bizkaiko Diputazioaren Jauregian.
4. Gu Gorbea aldeko bazter guztietatik ibiliak gara.
5. NASAk abenduaren 28an aireratu zuen globo batek 42 egun baino gehiago egin ditu Antartika gainean.

5 Euskararako beste tresna linguistiko batzuk

Morpheus-en analisi osoak
ZT Corpusa <http://www.ztcorpusa.net/>_
EDBL datu base lexikalaren kontsulta <http://ixa2.si.ehu.es/demo/edbl.jsp>
Entitateak

<http://ixa2.si.ehu.es/demo/entitateak.jsp>
Euskararako adiera desanbiguazioa <http://ixa3.si.ehu.es/wsd-demo/>
EuskalWordNet-en kontsularako interfazea <http://ixa2.si.ehu.es/mcr/wei.html>
EuSemcor-ren kontsultarako interfazea <http://sisx04.si.ehu.es:8080/eusemcor/>
Opentrad itzultzailea <http://www.opentrad.org/demo/?language=eu>

6 Gaztelania lantzen: Freeling eta Ancora.

6.1 Kategoria multzoa

Freeling-ek eta Ancora sistemek erabiltzen dituzten kategoriak PAROLE multzokoak dira, EAGLES taldeak definitu zuen bera. Freeling sistemaren web orrian ere ikus daiteke kategoria horien azalpen zabala *Documentation* atalean, konkretuki beste web orri honetan:

http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html

Parole etiketek sistema desberdin bat jarraitzen dute. Adibidez, *chico* hitzaren etiketa hau da: NCMS000; non NCMS letrek adierazten duten izen arrunta maskulino singular dela (NCMS: NombreComúnMasculinoSingular). Kategoria nagusi bakoitzaren ezaugarrien balioak letra-kode bereziekin adierazten dira; esaterako izenetan maskulino eta singular ezaugarria M eta S letrekin 3. eta 4. posizioetan. Kategoria nagusiak hauek dira:

- adjetivos (A)
- adverbios (R)
- artículos (T)
- determinantes (D)
- nombres (N)
- verbos (V)
- pronombres (P)
- conjunciones (C)
- interjecciones (I)
- preposiciones (S)
- signos de puntuación (F)
- numerales (Z)
- fechas y horas (W)

Gero kategoria nagusi horietako bakoitzak bere kode bereziak ditu ezaugarrien balioak adierazteko. Adibidez, izenentzako ezaugarriak, balioak eta kodeak honako hauek dira:

NOMBRES

Pos.	**Ezaugarria**	**Balioa**	**Kodea**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	-	0

6	Género Semántico	-	0
7	Grado	Apreciativo	A

Aditzenak, aldiz, hauek dira:

ADITZAK/ VERBOS

Pos.	**Ezaugarria**	**Balioa**	**Kodea**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Condicional	C
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

6.2 Ariketa:

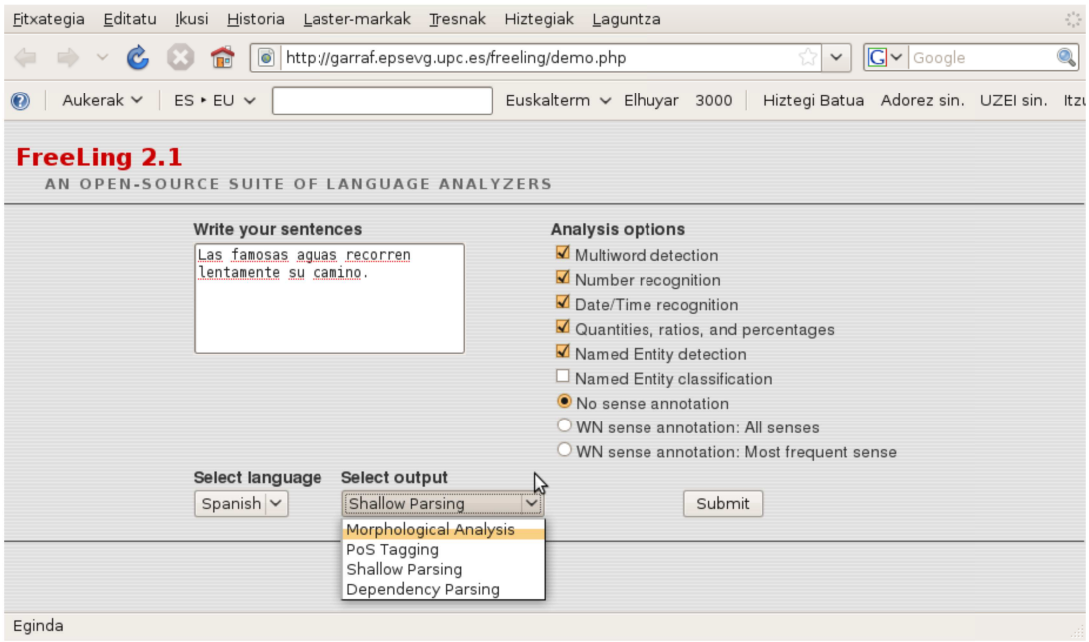
A. Ondoko zutabeak aztertu eta lotu hitz bakoitza bere etiketarekin, nahastuta daude:

cantas	NP000G0
éramos	NCMS00D
cantadas	NCMN000
cantarías	VMIC2S0
Barcelona	VMP00P***F***
gatito	VMIP2S0
cortapapeles	VSII1P0

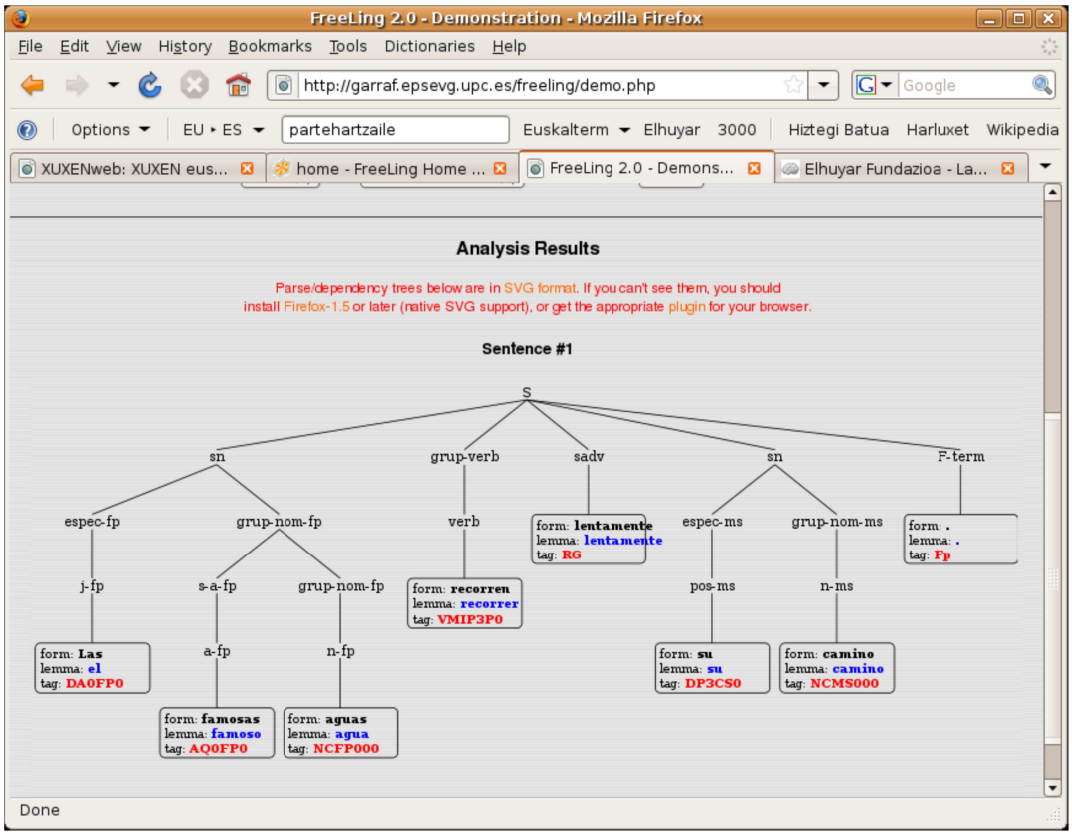
6.3 Freeling

Bilatu “Freeling” analizatzailea Interneten eta ezkerreko menuan aukeratu “on-line demo” sistema zuzenean nola dabilen ikusi ahal izateko. Helbide honetara ailegatuko zara: http://garraf.epsevg.upc.es/freeling/demo.php

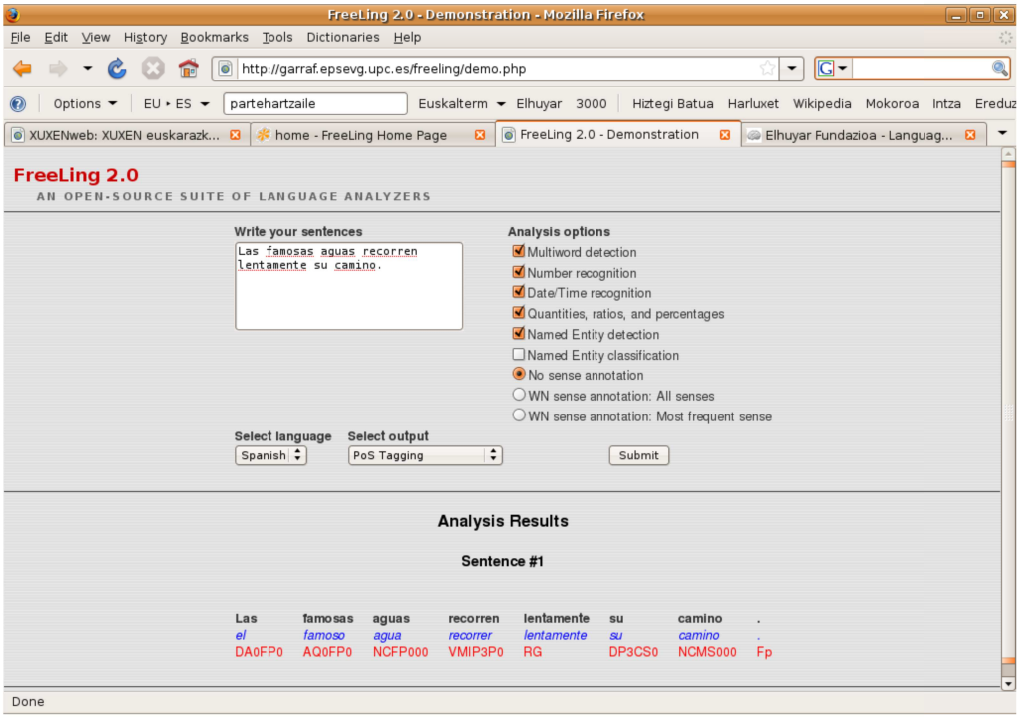
Hor zaudela aukeratu “Shallow parsing” (azaleko analisia) “select output” kutxan, eta analiza ezazu honako perpausa: “*Las famosas aguas recorren lentamente su camino.*”



Honako emaitza hau lortuko duzu:



Orain “select output” kutxan aukera ezazu “POS tagging” (etiketatu kategoriekin) “Shallow parsing” (azaleko analisia) izan ordez. Emaitzan hitz soilak ikusiko dituzu, ez zuhaitzik eta hitz bakoitzarekin bere kategoria:



6.4 Ancora

AnCora euskarazko, gaztelaniazko eta katalanezko corpusa da, hainbat hizkuntza-mailatan etiketatua. Beraz espainierazko perpausen analisi sintaktikoak ere ikus ditzakegu; bere erabilera euskararako definitu dugun erabilera bera da.

Beraz, besterik gabe jo ezazu zuzenean helbide honetara: http://clic.ub.edu/ancora/index.php. Edo lehen bezala, "IXA taldearen orritik(ixa.si.ehu.es) abiatuta, aukera ezazu *Demoak* ezkerreko menuan eta bertan aukeratu "Ancora: *zuhaitz sintaktikoen bankua*". Aldaketa bakarra da hasierako orrian goian eskuinaldean aztertu nahi den corpusa aukeratu behar dela: CESS_EU euskarazko esaldiak aztertzeko, AnCora_CA katalanerako eta AnCora_ES espainierarako.

Espainierazko esaldiak analizatzeko beraz AnCora_ES corpusa aukeratu bilatu aztertu nahi duzun esaldia eta ikus ezazu bere analisia modu grafikoan ala testu gisa.

7 Ingelesa lantzen. Freeling eta Conexor.

7.1 Kategoria multzoa.

Ingeleserako gehien erabiltzen den kategoria multzoa *Penn Treebank*-a osatzeko erabili zena da. Etiketa honen erreferentzia eta azalpen nagusia Beatriz Santorini-ren artikulua[1] izan daiteke edo sarean aurki daitezkeen beste webgune batzuk ere[2]. Hemen aurkezten ditugu hitz mailan etiketatzeko erabiltzen diren etiketak, etiketa guztiak eranskinetan ikus daitezke.

CC - Coordinating conjunction
CD - Cardinal number
DT - Determiner
EX - Existential there
FW - Foreign word
IN - Preposition or subordinating conjunction
JJ - Adjective
JJR - Adjective, comparative
JJS - Adjective, superlative
LS - List item marker
MD - Modal
NN - Noun, singular or mass
NNS - Noun, plural
NNP - Proper noun, singular
NNPS - Proper noun, plural
PDT - Predeterminer
POS - Possessive ending
PRP - Personal pronoun
PRP$ - Possessive pronoun (prolog version PRP-S)
RB - Adverb
RBR - Adverb, comparative
RBS - Adverb, superlative
RP - Particle
SYM - Symbol
TO - to
UH - Interjection
VB - Verb, base form
VBD - Verb, past tense
VBG - Verb, gerund or present participle
VBN - Verb, past participle
VBP - Verb, non-3rd person singular present
VBZ - Verb, 3rd person singular present
WDT - Wh-determiner
WP - Wh-pronoun

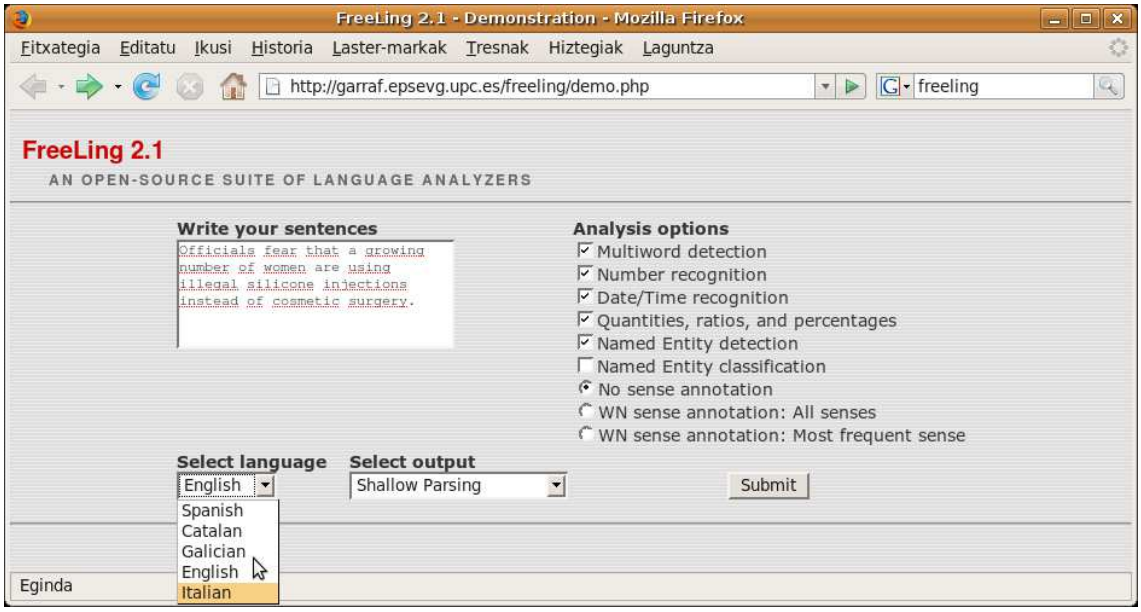
1 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps

2 http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html

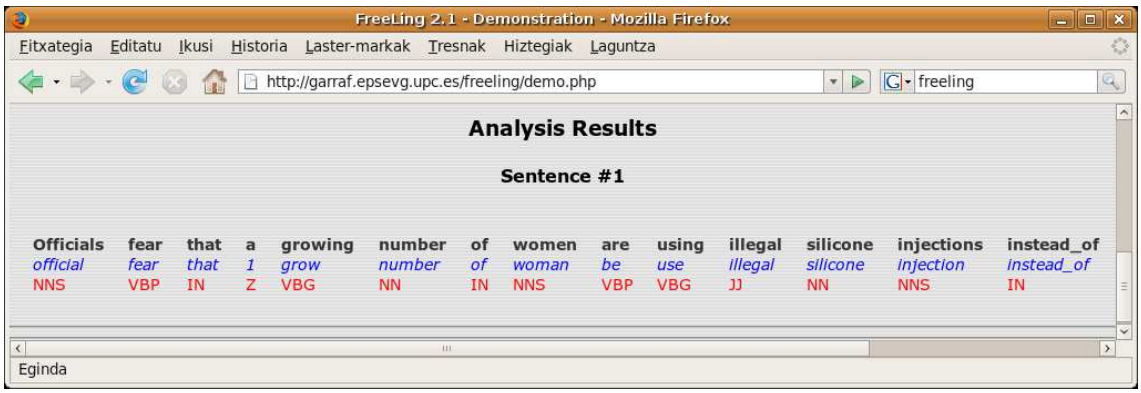
WP$ - Possessive wh-pronoun (prolog version WP-S)
WRB – Wh-adverb

7.2 Freeling

Freeling analizatzaile sintaktikoa lau hizkuntzatan erabil daiteke: espainiera, katalana, galiziera, ingelesa eta italiera. Beraz, espainierarako erakutsi dugun bezalaxe erabil daiteke ingeleserako ere, Nahikoa da hizkuntza aukeratzea beheko *Select Language* kutxan irudi honetan ikusten den bezala:



Ondoko irudi honetan ikus daiteke zer ematen digun POS_tagging aukeratuta (hitz bakoitzerako bere lema eta kategoria desanbiguatua). Lortu liteke zuhaitz osoa ere *Shalow parsing* aukera erabiliz.



8 Bibliografia

Oinarrizko bibliografia

Aduriz i., Aranzabe M.J., Arriola J.M. eta Díaz de Ilarraza A. 2006 **Sintaxi partziala.** In B. Fernández eta I. Laka (arg.) *Andolin gogoan: Essays in Honour of Professor Eguzkitza*. UPV/EHUko Argitarapen Zerbitzua, Bilbo.

Aduriz I. eta Díaz de Ilarraza A. 2003 **Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque.** In B. Oyharcabal (arg.), *Inquiries into the lexicon-syntax relations in Basque.* ASJUren gehigarria. UPV/EHU, Bilbo.

Aldezabal I., Arriola J., Díaz de Ilarraza A., Sarasola K. 2005
Hizkuntzalaritza Konputazionala
Liburuaren ISBN: 84-8438-065-3. HIZTEK saila. Udako Euskal Unibertsitatea

Alegria I., Miriam Urkia 2002
Morfologia Konputazionala. Euskararen morfologiaren deskribapena.
UEU. ISBN: 84-8438-034-3

Alegria i., Balza I., Ezeiza N., Fernández I., Urizar R. 2003 **Named Entity Recognition and Classification for texts in Basque.** II Jornadas de Tratamiento y Recuperación de Infomación. Madrid.

Alegria I., Artola X., Díaz de Ilarraza A., Sarasola K. 2008
Hizkuntza-teknologia Ixa taldean, euskararen erabilera errazteko eta sustatzeko aplikazioak
Bat Soziolinguistika. 66. zenb. 41-60 orr. ISSN:1130-8435

Ezeiza N. 1997. **EUSLEM, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa**. Tesina, Euskal Herriko Unibertsitatea

Freeling. User manual. http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.pdf

Freeling. Introducción a las etiquetas EAGLES (v 2.0)
http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html#cifras

Martí M.A., Taulé M., Bertran M., Màrquez L. **AnCora: Multilingual and Multilevel Annotated Corpora**

Santorini`B. **Part-of-Speech Tagging Guidelines for the Penn Treebank Project**.
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps

Bestelako bibliografia

Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. 2006 **Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing.** In Wilson A., Rayson P. eta Archer D. (arg.), *Corpus Linguistics Around the World*, 1-15. Rodopi (Netherland).

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. 2002
Robustness and customisation in an analyser/lemmatiser for Basque
LREC-2002 Customizing knowledge in NLP applications workshop, pages 1-6, Las Palmas de Gran Canaria, 28th May 2002

Palomar M., Civit M., Díaz de Ilarraza A., Moreno L., Bisbal E., Aranzabe M.J., Ageno A., Martí M.A., Navarro B. **3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano.** XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 81-88. Universidad de Barcelona, Barcelona.

9 Eranskinak

9.1 Euskararen kategoria multzoa

Kategoria Lexikalak (15)

Kategoria Nagusiak eta Azpikategoriak

- IZE izenak
  - ARR arruntak *(zuhaitz)*
  - IZB pertsona-izen bereziak *(Mikel)*
  - LIB leku-izen bereziak *(Donostia)*
  - ZKI zenbakia *(bat)*
- ADJ adjektiboak
  - ARR arruntak *(handi, benetako)*
  - GAL galdetzaileak *(nongo)*
- ADI aditzak
  - SIN sinpleak *(ekarri)*
  - ADK konposatuak *(lo egin)*
  - ADP perifrastikoak *(ahal izan)*
  - FAK faktitiboak (*etorrarazi*)
- ADB adberbioak
  - ARR arruntak *(gaur, negarrez)*
  - GAL galdetzaileak *(noiz)*
- DET determinatzaileak
  - ERK erakusleak
  - ERK ARR arruntak *(hau)*
  - ERKIND indartuak *(berori)*
  - NOL nolakotzaileak
  - NOLARR arruntak *(edozein)*
  - NOLGAL galdetzaileak *(zein)*
  - ZNB zenbatzaileak
  - DZH zehaztuak (*bi*)
  - BAN banatzaileak (*bina*)
  - ORD ordinalak (*bigarren*)
  - DZG zehaztugabeak *(zenbait)*
  - ORO orokorrak *(guzti)*
- IOR izenordainak
  - PER pertsonalak
  - PERARR arruntak *(ni)*
  - PERIND indartuak *(neu)*
  - IZG zehaztugabeak
  - IZGMGB mugagabeak *(norbait)*
  - IZGGAL galdetzaileak (*nor*)
  - BIH bihurkariak *(-(r)en burua)*
  - ELK elkarkariak *(elkar)*

- LOT loturazkoak
  - LOK lokailuak (*hala ere*)
  - JNT juntagailuak (*edo*)
- PRT partikulak (*omen, ote, ...*)
- ITJ interjekzioak (*alajaina!*)
- BST bestelakok (*baldin*)

Kategoria lagungarriak (5)

- ADL ADITZ LAGUNTZAILEAK (*du*)
- ADT ADITZ SINTETIKOAK (*dator*)
- SIG SIGLAK (*EHU*)
- SNB SINBOLOAK (*km, cm, g,...*)
- LAB LABURDURAK (*etab.*)

Kategoria Morfologikoak (9)

- AMM ADITZ-MOTA MORFEMAK *(-tu, -t(z)e,...)*
- ASP ASPEKTU-MORFEMAK (*Ø, -ko,...*)
- ATZ ATZIZKIAK (*-pe*)
- AUR AURRIZKIAK (*ber-*)
- DEK DEKLINABIDE MORFEMAK (*-aren*)
- ELI ELIPSIA (*Ø*)
- ERL ERLAZIO ATZIZKIAK (*-(e)la*)
- GRA GRADUATZAILEAK (*-ago*)
- MAR MARRA (-)

Puntuazio-zeinuak (3)

- PNT PUNTUA
- BPM BESTE PUNTUAZIO ZEINUAK *(puntuaren pareko izan daitezkeenak)*
- PSB PUNTUAZIO SINBOLOAK *(parentesiak, marra luzea, kakotxak,...)*

9.2 Gaztelaniaren kategoria multzoa

Parole etiketa multzoa EAGLES taldeak definitu zuen. Freeling sistemaren web orrian ere ikus daiteke kategoria horien azalpen zabala *Documentation* atalean[3]. Hauek dira kategoria nagusien kode bereziak ezaugarrien balioak adierazteko.

ADJETIVOS

Pos.	Atributo	Valor	Código
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
3	Grado	Apreciativo	A
4	Género	Masculino	M
		Femenino	F
		Común	C
5		Número	Singular
			Plural
			Invariable
6	Caso	-	0
7	Función	Participio	P

ADVERBIOS

Pos.	Atributo	Valor	Código
1	Categoría	Adverbio	R
2	Tipo	General	G

ARTÍCULOS

Pos.	Atributo	Valor	Código
1	Categoría	Artículo	T
2	Tipo	Definido	D
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
5	Caso	-	0

DETERMINANTES

Pos.	Atributo	Valor	Código
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
3	Persona	Primera	1
		Segunda	2

[3] http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html

		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Caso	-	0
7	Poseedor	1ª persona-sg	1
		2ª persona-sg	2
		3ª persona	0
		1ª persona-pl	4
		2ª persona-pl	5

NOMBRES

Pos.	Atributo	Valor	Código	
1	Categoría	Nombre	N	
2	Tipo	Común	C	
		Propio	P	
3	Género	Masculino	M	
		Femenino	F	
		Común	C	
4	Número	Singular	S	
		Plural	P	
		Invariable	N	
5	Caso	-	0	
6	Género Semántico		-	0
7	Grado	Apreciativo	A	

VERBOS

Pos.	Atributo	Valor	Código
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Condicional	C
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P

7	Género	Masculino	M
		Femenino	F

PRONOMBRES

Pos.	Atributo	Valor	Código
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Caso	Nominativo	N
		Acusativo	A
	Dativo	D	
		Oblicuo	O
7	Poseedor	1ª persona-sg	1
		2ª persona-sg	2
		3ª persona	0
		1ª persona-pl	4
		2ª persona-pl	5
8	Politeness	Polite	P

CONJUNCIONES

Pos.	Atributo	Valor	Código
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S
3	-	-	0
4	-	-	0

NUMERALES

Pos.	Atributo	Valor	Código
1	Categoría	Numeral	M
2	Tipo	Cardinal	C
		Ordinal	O
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P

5	Caso	-	0
6	Función	Pronominal	P
		Determinante	D
		Adjetivo	A

INTERJECCIONES

Pos.	Atributo	Valor	Código
1	Categoría	Interjección	I

ABREVIATURAS

Pos.	Atributo	Valor	Código
1	Categoría	Abreviatura	Y

PREPOSICIONES

Pos.	Atributo	Valor	Código
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
3	Género	Masculino	M
4	Número	Singular	S

SIGNOS DE PUNTUACIÓN

Pos.	Atributo	Valor	Código
1	Categoría	Puntuación	F

9.3 Ingelesaren kategoria multzoa

9.3.1 Clause Level

S - simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a wh-word and that does not exhibit subject-verb inversion.

SBAR - Clause introduced by a (possibly empty) subordinating conjunction.

SBARQ - Direct question introduced by a wh-word or a wh-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.

SINV - Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.

SQ - Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ.

9.3.2 Phrase Level

ADJP - Adjective Phrase.

ADVP - Adverb Phrase.

CONJP - Conjunction Phrase.

FRAG - Fragment.

INTJ - Interjection. Corresponds approximately to the part-of-speech tag UH.

LST - List marker. Includes surrounding punctuation.

NAC - Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.

NP - Noun Phrase.

NX - Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.

PP - Prepositional Phrase.

PRN - Parenthetical.

PRT - Particle. Category for words that should be tagged RP.

QP - Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.

RRC - Reduced Relative Clause.

UCP - Unlike Coordinated Phrase.

P - Vereb Phrase.

WHADJP - Wh-adjective Phrase. Adjectival phrase containing a wh-adverb, as in how hot.

WHAVP - Wh-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a wh-adverb such as how or why.

WHNP - Wh-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some wh-word, e.g. who, which book, whose daughter, none of which, or how many leopards.

WHPP - Wh-prepositional Phrase. Prepositional phrase containing a wh-noun phrase (such as of which or by whose authority) that either introduces a PP gap or is contained by a WHNP.

X - Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing the...the-constructions.

9.3.3 Word level

CC - Coordinating conjunction
CD - Cardinal number
DT - Determiner
EX - Existential there
FW - Foreign word
IN - Preposition or subordinating conjunction
JJ - Adjective
JJR - Adjective, comparative
JJS - Adjective, superlative
LS - List item marker
MD - Modal
NN - Noun, singular or mass
NNS - Noun, plural
NNP - Proper noun, singular
NNPS - Proper noun, plural
PDT - Predeterminer
POS - Possessive ending
PRP - Personal pronoun
PRP$ - Possessive pronoun (prolog version PRP-S)
RB - Adverb
RBR - Adverb, comparative
RBS - Adverb, superlative
RP - Particle
SYM - Symbol
TO - to
UH - Interjection
VB - Verb, base form
VBD - Verb, past tense
VBG - Verb, gerund or present participle
VBN - Verb, past participle
VBP - Verb, non-3rd person singular present
VBZ - Verb, 3rd person singular present
WDT - Wh-determiner
WP - Wh-pronoun
WP$ - Possessive wh-pronoun (prolog version WP-S)
WRB - Wh-adverb

9.3.4 Function tags. Form/function discrepancies

-ADV (adverbial) - marks a constituent other than ADVP or PP when it is used adverbially (e.g. NPs or free ("headless" relatives). However, constituents that themselves are modifying an ADVP generally do not get -ADV. If a more specific tag is available (for example, -TMP) then it is used alone and -ADV is implied. See the Adverbials section.

-NOM (nominal) - marks free ("headless") relatives and gerunds when they act nominally.

9.3.5 Function tags. Grammatical role

-DTV (dative) - marks the dative object in the unshifted form of the double object construction. If the preposition introducing the "dative" object is for, it is considered benefactive (-BNF). -DTV (and -BNF) is only used after verbs that can undergo dative shift.

-LGS (logical subject) - is used to mark the logical subject in passives. It attaches to the NP object of by and not to the PP node itself.

-PRD (predicate) - marks any predicate that is not VP. In the do so construction, the so is annotated as a predicate.

-PUT - marks the locative complement of put.

-SBJ (surface subject) - marks the structural surface subject of both matrix and embedded clauses, including those with null subjects.

-TPC ("topicalized") - marks elements that appear before the subject in a declarative sentence, but in two cases only:

1. if the front element is associated with a *T* in the position of the gap.
2. if the fronted element is left-dislocated (i.e. it is associated with a resumptive pronoun in the position of the gap).

-VOC (vocative) - marks nouns of address, regardless of their position in the sentence. It is not coindexed to the subject and not get -TPC when it is sentence-initial.

9.3.6 Function tags. Adverbials

Adverbials are generally VP adjuncts.

-BNF (benefactive) - marks the beneficiary of an action (attaches to NP or PP).

This tag is used only when (1) the verb can undergo dative shift and (2) the prepositional variant (with the same meaning) uses for. The prepositional objects of dative-shifting verbs with other prepositions than for (such as to or of) are annotated -DTV.

-DIR (direction) - marks adverbials that answer the questions "from where?" and "to where?" It implies motion, which can be metaphorical as in "...rose 5 pts. to 57-1/2" or "increased 70% to 5.8 billion yen" -DIR is most often used with verbs of motion/transit and financial verbs.

-EXT (extent) - marks adverbial phrases that describe the spatial extent of an activity. -EXT was incorporated primarily for cases of movement in financial space, but is also used in analogous situations elsewhere. Obligatory complements do not receive -EXT. Words such as fully and completely are absolutes and do not receive -EXT.

-LOC (locative) - marks adverbials that indicate place/setting of the event. -LOC may also indicate metaphorical location. There is likely to be some varation in the use of -LOC due to differing annotator interpretations. In cases where the annotator is faced with a choice between -LOC or -TMP, the default is -LOC. In cases involving SBAR, SBAR should not receive -LOC. -LOC has some uses that are not adverbial, such as with place names that are adjoined to other NPs and NAC-LOC premodifiers of NPs. The special tag -PUT is used for the locative argument of put.

-MNR (manner) - marks adverbials that indicate manner, including instrument phrases.

-PRP (purpose or reason) - marks purpose or reason clauses and PPs.

-TMP (temporal) - marks temporal or aspectual adverbials that answer the questions when, how often, or how long. It has some uses that are not

strictly adverbial, auch as with dates that modify other NPs at S- or VP-level. In cases of apposition involving SBAR, the SBAR should not be labeled -TMP. Only in "financialspeak," and only when the dominating PP is a PP-DIR, may temporal modifiers be put at PP object level. Note that -TMP is not used in possessive phrases.

9.3.7 Function tags. Miscellaneous

-CLR (closely related) - marks constituents that occupy some middle ground between arguments and adjunct of the verb phrase. These roughly correspond to "predication adjuncts", prepositional ditransitives, and some "phrasel verbs". Although constituents marked with -CLR are not strictly speaking complements, they are treated as complements whenever it makes a bracketing difference. The precise meaning of -CLR depends somewhat on the category of the phrase.

* on S or SBAR - These categories are usually arguments, so the -CLR tag indicates that the clause is more adverbial than normal clausal arguments. The most common case is the infinitival semi-complement of use, but there are a variety of other cases.

* on PP, ADVP, SBAR-PRP, etc - On categories that are ordinarily interpreted as (adjunct) adverbials, -CLR indicates a somewhat closer relationship to the verb. For example:

o Prepositional Ditransitives

In order to ensure consistency, the Treebank recognizes only a limited class of verbs that take more than one complement (-DTV and -PUT and Small Clauses) Verbs that fall outside these classes (including most of the prepositional ditransitive verbs in class [D2]) are often associated with -CLR.

o Phrasal verbs

Phrasal verbs are also annotated with -CLR or a combination of -PRT and PP-CLR. Words that are considered borderline between particle and adverb are often bracketed with ADVP-CLR.

o Predication Adjuncts

Many of Quirk's predication adjuncts are annotated with -CLR.

* on NP - To the extent that -CLR is used on NPs, it indicates that the NP is part of some kind of "fixed phrase" or expression, such as take care of. Variation is more likely for NPs than for other uses of -CLR.

-CLF (cleft) - marks it-clefts ("true clefts") and may be added to the labels S, SINV, or SQ.

-HLN (headline) - marks headlines and datelines. Note that headlines and datelines always constitute a unit of text that is structurally independent from the following sentence.

-TTL (title) - is attached to the top node of a title when this title appears inside running text. -TTL implies -NOM. The internal structure of the title is bracketed as usual.