EUSEMCOR:
euskarako corpusa semantikoki etiketatzeko eskuliburua;
editatze-, etiketatze- eta epaitze-lanak

Agirre, E.; Aldezabal, I.; Etxeberria, J.; Izagirre, E.; Mendizabal, K.; Pociello, E.; Quintian, M.

2005-10-14

Laburpena

IXA taldean semantikoki etiketatutako corpusa garatzen ari gara euskarako. Txosten honen helburua corpus etiketatu hori eraikitzen lagunduko digun eskuliburu bat egitea izan da.

Abstract

IXA Research group is developing a semantically tagged corpus for Basque. The aim of this report is to build a guide in which we can find all the required information for this task.

AURKIBIDEA:

1. Lanaren helburuak

Lan honen helburua 300.000 hitzeko corpusa (EuSemCor[1] deritzoguna) euskaraz etiketatzea da EuskalWordNet[2]-eko adierak edo synset-ak[3] erabiliaz. Proiektu honetan izenak, adjektiboak eta aditzak etiketatuko dira. Aldi berean, eta corpusetik lortzen den informazioan oinarrituz, EuskalWordNet-eko synset-ak orraztuko dira; hau da, behin 300.000 hitzeko corpusaren etiketatze semantikoa amaituta, EuskalWordNet-ek corpusean agertu diren adiera horiek guztiak izan beharko ditu.

2. Metodologia

Puntu honetan EuSemCor-en etiketatze semantikoan erabili den lan metodologia azalduko dugu.

2.1. Lan-taldea

- 2 etiketatzaile/hizkuntzalari (egun erdia) → **Eli Izagirre eta Mikel Quintian**
- 1 epaile / hizkuntzalari (egun osoa) → **Jone Etxeberria (Izaskun Aldezabal)**
- 1 editore / hizkuntzalari (egun erdia) → **Karmele Mendizabal (Eli Pociello)**
- 1 informatikari → **Kike Fernández (Eneko Agirre)**
- 1 hizkuntzalarien koordinatzaile → **Izaskun Aldezabal**
- 1 lexikografo? → **Elhuyar**

2.2. Lan-taldearen funtzionamendu orokorra

Corpuseko hitzak zerrendatuak daude corpusean duten maiztasunaren arabera (maiztasun handienetik txikienera). Zerrenda honetatik **editoreak** aukeratzen ditu landu beharreko hitzak[4], eta hitz hauen EuskalWordNet-eko synset-ak orraztuko ditu[5]. Hitzak orraztu ondoren, **editorea, etiketatzaileak eta epailea** elkartuko dira hitz horien synset-en esanahia ulertzeko.

---

[1] Etiketatu beharreko corpusa *Euskaldunon Egunkaria*ko berriekin eta *XX. mendeko Euskararen Corpus Estatistikoa*rekin osatua dago.

[2] EuskalWordNet zer den A.2.1 atalean azaltzen da. Bestalde, EuskalWordNet-en 1.6 bertsioa erabiltzen dugula azpimarratu beharra dago (http://siuc02.si.ehu.es/wei2004-06-21/wei.html).

[3] Aurrerantzean, **EuskalWordNet-eko adierak** adierazteko *synset* terminoa erabiliko dugu. *Adiera* termino orokorragoa da, eta hau erabiliko dugu EuskalWordNet-en ez dauden beste kontzeptu horiek adierazteko (hiztegietan eta corpusean agertzen diren kontzeptu berriak, batez ere).

[4] Berez, zerrenda hau bitan banatuta dago: zerrenda batek corpusean usuenak diren hitzak izango ditu (`landutakoak.n.usuenak.txt`), eta besteak, berriz, ez usuak direnak (`landutakoak.n.ezusuak.txt`). ~jirhizts/Corpus/PROFIT2/koordinazioa fitxategian (hauetara iristeko azalpena 4.1 atalean dago). Hasiera batean, bi zerrendetatik hartutako izenak lantzen baziren ere, gaur egun, usuen zerrenda bakarrik baliatzen da, hau bukatu arte behintzat.

[5] Editorearen lanaren berri izateko, ikus 4. atala eta A eranskina.

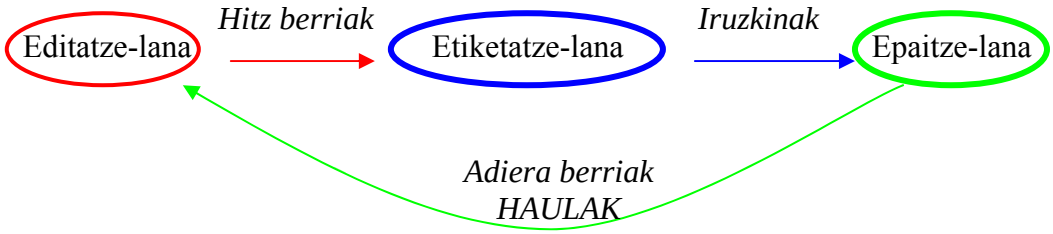
Editoreak, epaileak eta etiketatzaileek hitzen synset-ak zeintzuk diren ulertu eta adostu dutenean, **etiketatzaileak** hitzei dagozkien agerpenak etiketatzen hasiko dira[6].

Etiketatu beharrekoak bukatu ondoren, **etiketatzaileek editoreari** eta **epaileari** jakinaraziko diete eta etiketatzean izan dituzten gorabeherak azalduko dituzte bilera batean[7].

Gero, **epaileak** bi etiketatzaileen lana erkatuko du eta ezberdin etiketatuta dauden agerpen horiek ebatziko ditu. Bestalde, aldaketaren bat egingo balu glosen itzulpenak eta adibideak begiratu beharko lituzke. Eta azkenik, corpusean agertu diren adiera berriak jakinarazi beharko dizkio **editoreari**[8]**.**

Editoreak corpusean agerturiko adiera berri horien egokitasuna aztertuko du hauek EuskalWordNet-en sartzea erabaki baino lehen.

Ikus daitekeen bezala, metodologia ziklikoa da:



2.3. Bilerak

Hasiera batean, astean behin bilera bat egingo litzateke, eta bileretan zera egingo da:

a. **Editoreak** hitz berriak aurkeztu eta banatuko ditu; hauetan zalantzak izan baditu, non eta zergatik azalduko du.

b. **Etiketatzaileek:**
   - Etiketatu berri dituzten hitzen gorabeherak komentatuko dituzte (zalantzak etiketatzean, adiera berriak, HAULak eta abar).
   - Editoreak banatutako hitzen synset-ak ulertzen/adosten saiatuko dira.

c. **Epaileak:**
   - Etiketatutako azken hitzen emaitzak aurrean dituela, etiketatzaileen zalantzak eta iruzkinak[9] jaso eta epaitzeko lagungarri izango zaion informazioa bilduko ditu[10].
   - Aurreko astean epaitutako hitzei buruzko emaitzak/zalantzak/erabakiak jakinaraziko ditu.

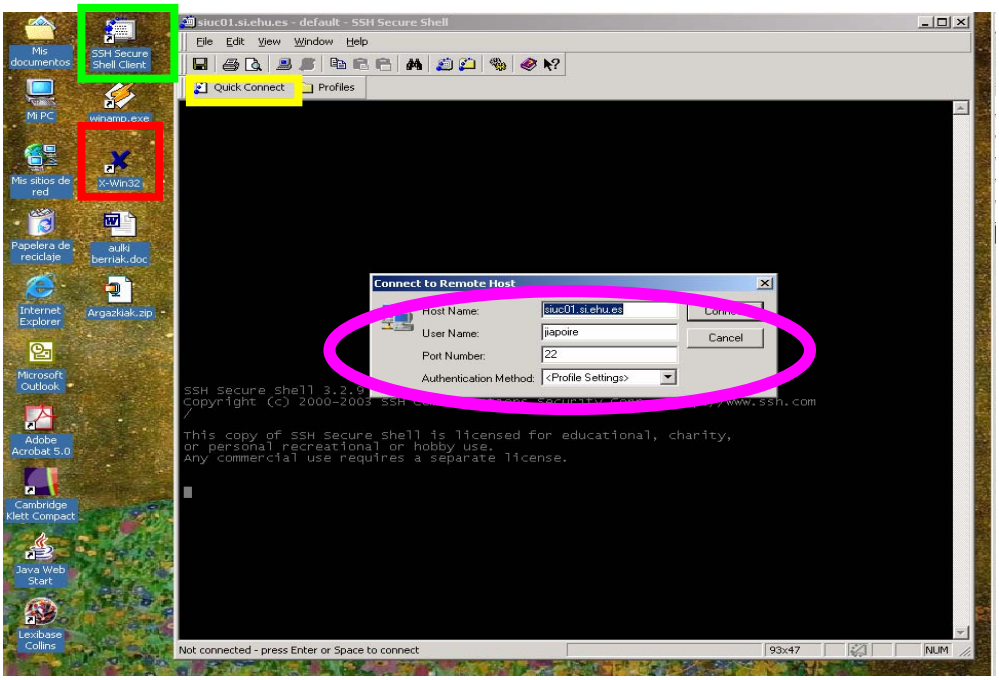
---

[6] Etiketatzaileen lanaren berri izateko, ikus 5. atala.
[7] Bileren berri izateko ikus 2.3 atala.
[8] Epailearen lanaren berri izateko, ikus 6. atala.
[9] Iruzkinen azkeneko bertsioa IxAko web orrian egon behar du eskuragarri: http://ixa.si.ehu.es/Ixa/Azpitaldeak/Lexikoa%20eta%20semantika/dokumentuak/Eskuzko%20etiketatze%20automatikoaren%20iruzkinetarako%20taula
[10] Bileran igarotako denbora ere apuntatu behar da.

3. 3lb tresna eta bere erabilerak

Sarreran aipatu bezala, proiektu honen helburua semantikoki etiketatutako corpusa eraikitzea da, eta horretarako, 3lb tresna erabiltzen da. Tresna honen bidez, corpusa etiketatzeaz gain, corpusean agertzen diren hitzei buruzko kontsultak ere egin daitezke.

EuSemCor corpusa fitxategi ezberdinetan banatua dago, fisikoki siuc01 edo siuc02 makinetan daude, eta fitxategi hauek 3lb tresnarekin irekitzen dira. Horretarako, *SSH Secure Shell Client* programaren exekutagarria edo honen mahai-gaineko lasterbidea sakatu (ikus 1go irudian karratu berdez aukeratua dagoen ikonoa) behar da. Kontuan izan, *SSH Secure Shell Client* erabili ahal izateko beharrezkoa dela *X-win* aplikazioa martxan egotea (ikus 1go irudian karratu gorriz aukeratua dagoen ikonoa). Behin *SSH Secure Shell Client* programaren interfazean gaudela *Quick Connect* sakatu behar da (ikus 1go irudian karratu horiz aukeratua dagoena), eta bertan *Host name* eremua (siuc01.si.ehu.es edo siuc02.si.ehu.es) eta *User name* eremua bete beharko dira (ikus 1go irudia). Ondoren, konektatzeko sakatu *Connect* eta pasahitza eskatuko du:



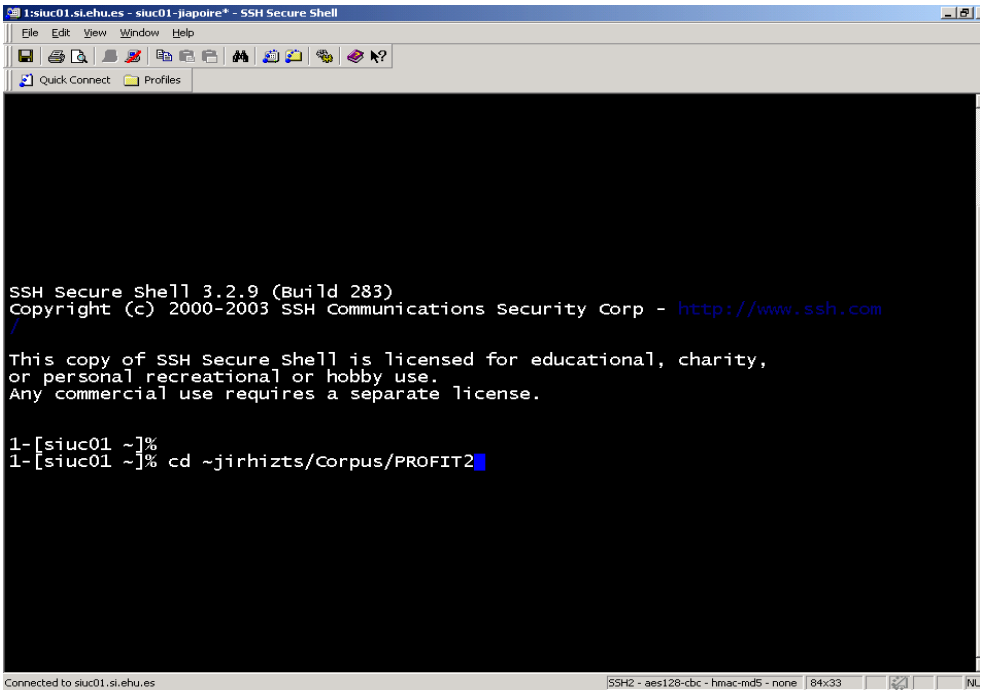
1go irudia

Datu horiek guztiak bete ondoren, interfazeak hurrengo itxura izango du (ikus 2. irudia)

```
1-[siuc01 ~]
```

3lb-rekin lanean hasteko, ondorengo direktorioan kokatu behar da (ikus 2. irudia):

```
2-[siuc01 ~]% cd ~jirhizts/Corpus/PROFIT2
```



2. irudia

3.1. Fitxategiak bilatu

Etiketatu beharreko hitzaren agerpen guztiak **zein fitxategitan dauden jakiteko**, direktorio honetan programa bat exekutatu behar da. Demagun *gezi* hitza landu nahi dela corpusean; horretarako `./bilatu `hitza´` beheko agindua idatzi beharko litzateke (kasu honetan landu nahi den hitza *gezi* denez hauxe idatziko da):

```
3-[siuc01 PROFIT2]% ./bilatu gezi
```

Galdeketa honen emaitza hau izango da:

```
1 eebs.450520392.sat.xml
1 eebs.4715023603.sat.xml
--------------------
Guztira:     2
```

Adibide honetan zera egin da: programari *gezi* hitza agertzen den fitxategi guztiak aurkitzeko agindua eman. Hitz hau bi fitxategitan dago (`eebs.450520392.sat.xml` eta `eebs.4715023603.sat.xml`), eta hauen aurrean ikus daitezkeen zenbakiek *gezi* hitzak fitxategi bakoitzean duen agerpen kopurua erakusten dute. Bistan denez, kasu honetan, fitxategi bakoitzean agerpen bakarra etiketatu beharko da.

Lema batzuk kategoria bat baino gehiago izan dezakete. Kasu horietan fitxategien bilaketa gehiago zehatz dezakegu, bilatu nahi den hitzaren kategoria gehituz. Adibidez, lema izen bat denean (*gezi*) "n" bat idatziko da:

```
4-[siuc01 PROFIT2]%./bilatu gezi n
```

Aditz bat denean (*etorri*) “v” bat:

```
5-[siuc01 PROFIT2]%./bilatu etorri v
```

Eta adjektibo bat denean (*gorri*) “a” bat:

```
6-[siuc01 PROFIT2]%./bilatu gorri a
```

./`bilatu` komandoa etiketatzen hasi baino lehen erabiltzea gomendagarria da, horrela, etiketatzaileak hasiera batetik daki zeintzuk diren etiketatu behar dituen fitxategiak. Gainera, ./`bilatu` komando honen emaitza inprimatzea ere ondo etortzen da. Horretarako, komando hau behar da:

```
7-[siuc01 PROFIT2]% bilatu gezi | a2ps
```

Bestela, *SSH Secure Shell Client*-aren interfazeak berak inprimatzeko duen botoia ere erabil daiteke.

3.2. Fitxategiak ireki

Etiketatu beharreko fitxategiak zeintzuk diren jakinda, hauek bi bide erabilita ikus daitezke 3lb tresnan: batean, fitxategiak 3lb interfazea bera erabilita irekitzen dira, eta bestean, *SSH Secure Shell Client*-aren interfazetik bertatik `./ireki` komandoaren bitartez. Nolanahi ere, bigarren aukera askozaz ere erabilgarriagoa da etiketatu beharreko hitza zuzenean irekitzen baitu[11].

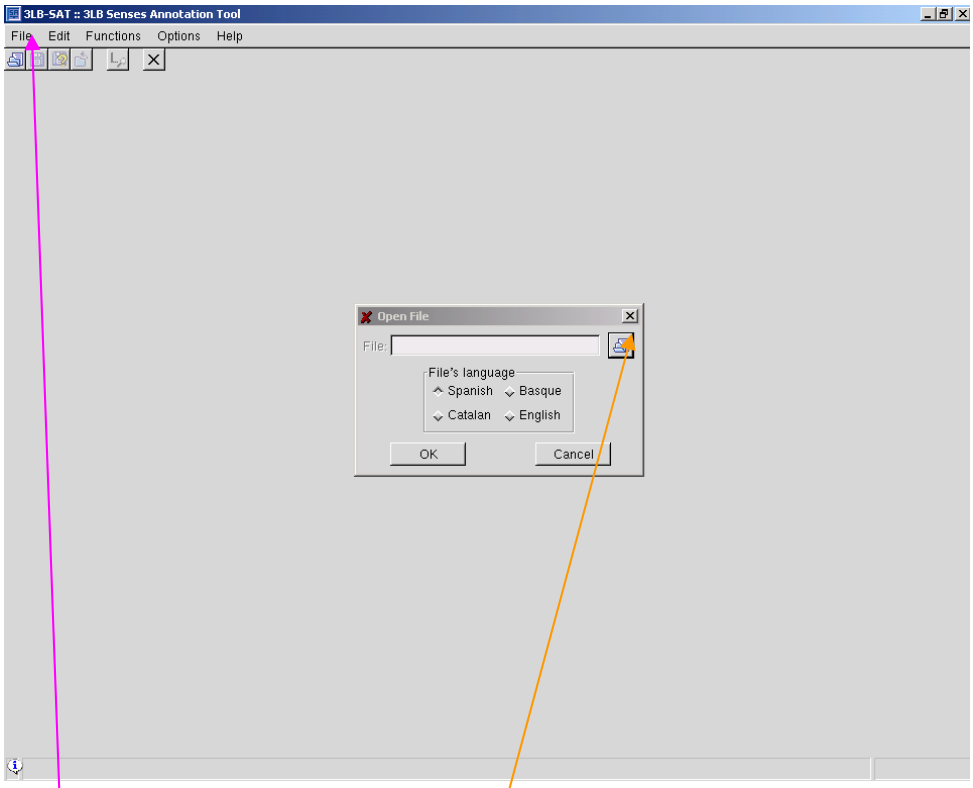
3.2.1. Fitxategiak ireki 3lb interfazea erabilita

Behin etiketatu beharreko fitxategiak ezagututa, 3lb tresnaren interfazea egikarituko da, beheko agindua emanaz:

```
8-[siuc01 PROFIT2]% 3lb-wn16 &
```

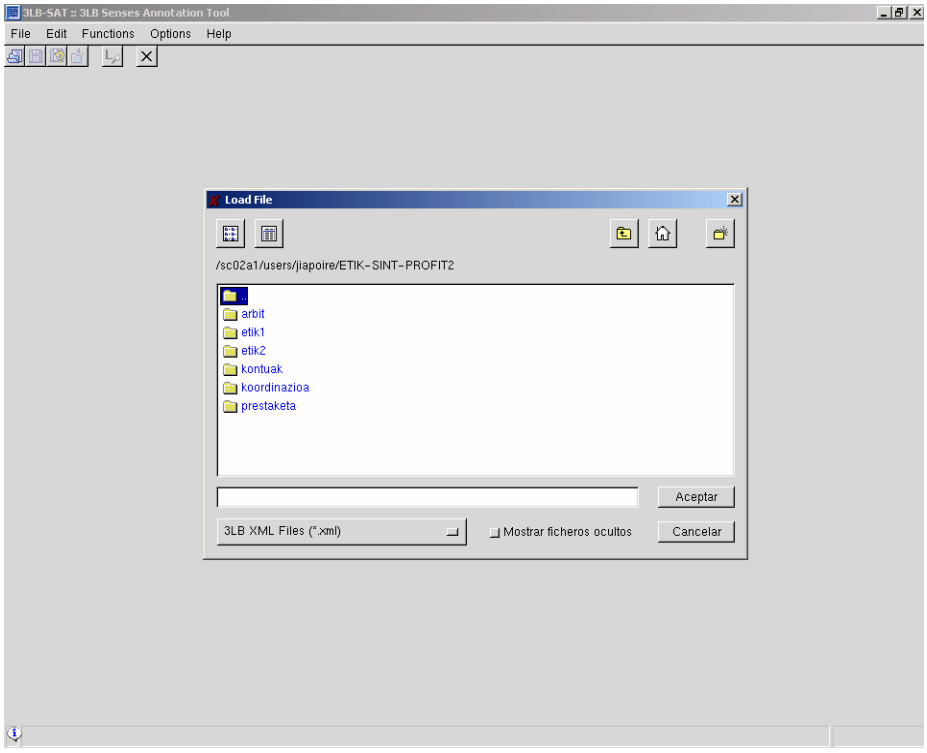
eta interfazean ondorengo pausoak jarraituko dira:

[11] Proiektu honen hasieran, 3lb interfazea erabiltzeko aukera bakarra interfazea bera erabilita zen. Gerora, interfazea erabiltzeak etiketatze-lana moteldu egiten zuenez, `./ireki` komandoa sortu zen. Txosten honetan, bi erabilera posibleak aipatzen dira, lanerako ditugun aukera guztiak ezagutzea gomendagarria delako. Hala ere, gaur egun erabiltzen dena bigarren aukera da, `./ireki` komandoa, alegia.



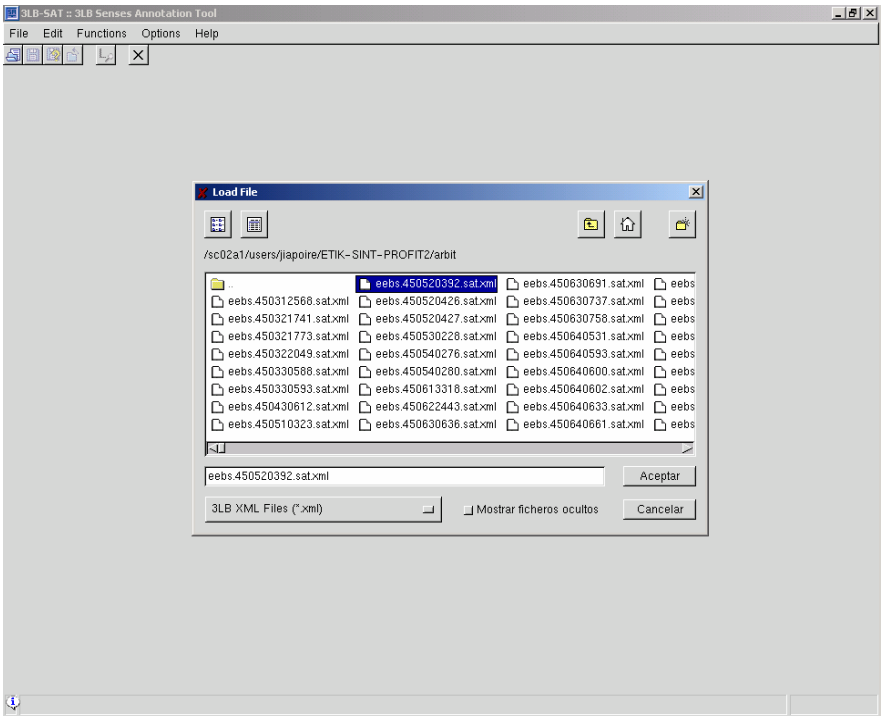
3. irudia

`File` eta `Load` botoiari eman ondoren, goiko irudian dugun pantaila agertuko da, eta bertan eskuineko laukitxoan egingo da klik (3. irudia).



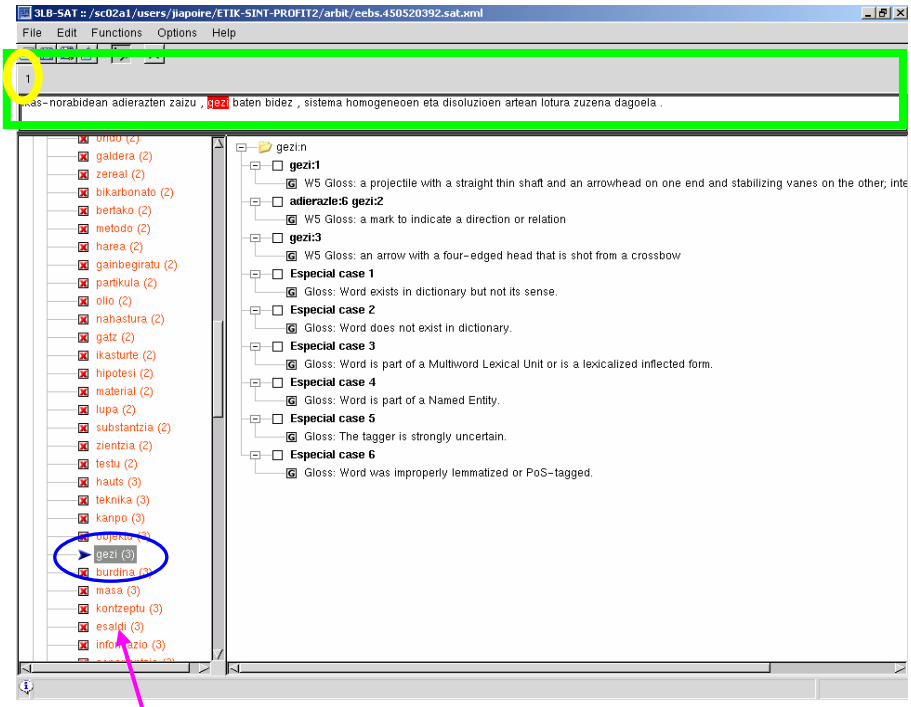
4. irudia

Agertuko den pantailan, **`ETIK-SINT-PROFIT2`** karpeta aukeratu lehendabizi, eta **`epai`** **(**epaileak**), `etik1` (“**etiketatzaile 1”-ek**)** edo **`etik2` (“**etiketatzaile 2”-k) karpetak ondoren (ikus 4. irudia). Orduan, etiketatu behar den fitxategia aukeratuko da (*gezi*ren adibidean `eebs.450520392.sat.xml` edo `eebs.4715023603.sat.xml`):



5. irudia

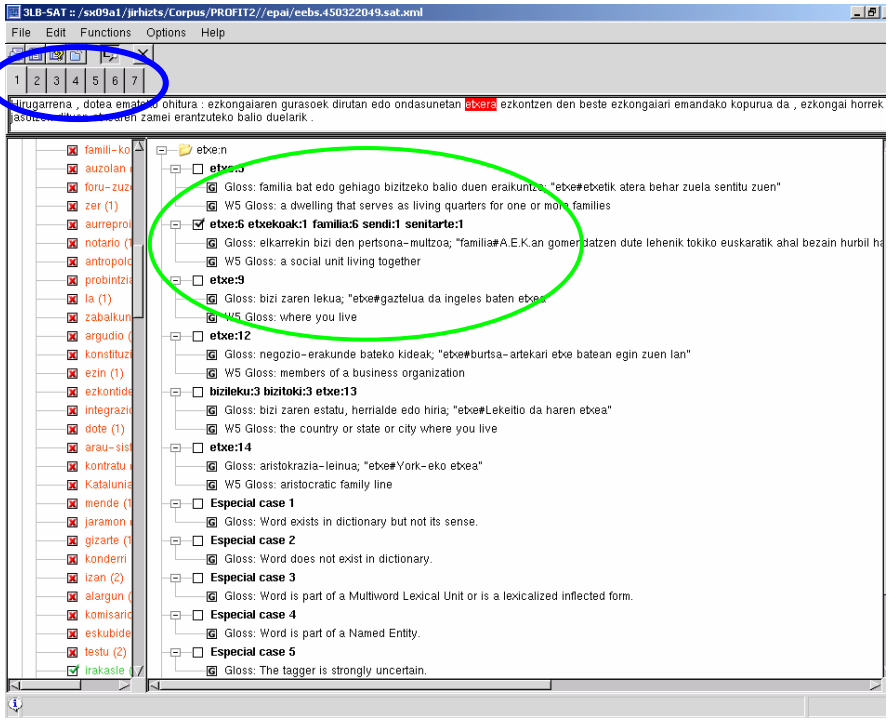
Eta ondoren, fitxategi horretan etiketatu behar den lema (*gezi*):



6. irudia

Lema-zerrendan, berdez dauden hitzak dagoeneko etiketatuak dauden hitzak dira. Gorriz daudenak, berriz, etiketatu beharrekoak. Hitzen bat horiz badago, hitz horrek fitxategi horretan dituen agerpenetako bat etiketatu gabe dagoela adierazten du.

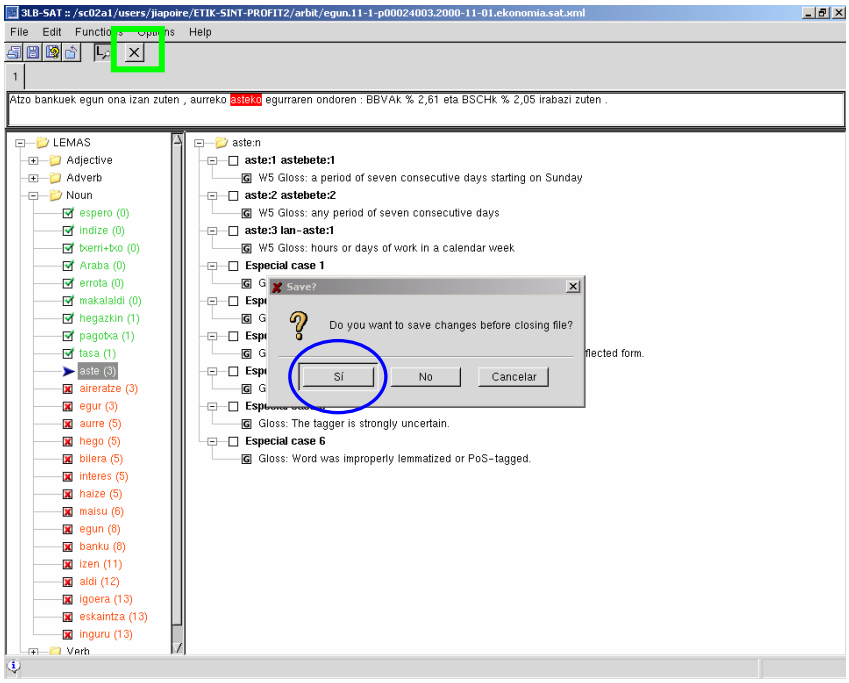
Interfazearen goiko aldean (ikus 6. irudiko karratu berdea), etiketatu beharreko agerpena (gorriz azpimarratua) eta honen testuinguruak azaltzen dira. 6. irudian agerpen bakarra dagoen arren (ikus borobil horia), fitxategietan agerpen bat baino gehiago egon daitezke, eta horiek zenbakituak datoz. Agerpen horiek ikusi/etiketatu ahal izateko, zenbaki horien gainean sakatu behar da (ikus 7. irudia).



7. irudia

Interfazearen eskuinaldean, EuskalWordNet-eko synset-zerrenda dago, eta horren gainean klikatu behar da agerpen bat etiketatzeko (ikus 7. irudian borobil berdea).

Fitxategi bateko etiketatze lana burutzean, beste hitz bat etiketatzeko eta pantaila horretatik ateratzeko `File` botoiari eman eta ondoren `Close`-ri eman; edo laburrago dena, `Options` botoiaren azpiko ixa-ri eman (ikus 8. irudian karratu berdez aukeratua dagoen ikonoa). Orduan, tresnak lana gorde nahi den galdetzen du. **PAUSO HAU OSO GARRANTZITSUA DA; zeren eta, `Sí` botoiari emango ez balitzaio, egindako lan guztia grabatu gabe gelditzeko litzateke** (ikus 8. irudia)**.**



8. irudia

3.2.2. Fitxategiak ireki `./ireki` komandoa erabilita

Agindu hau da 3lb-ra sartzeko modurik azkarrena, **behar diren fitxategi horiek guzitak *SSH Secure Shell Client*-aren interfazetik bertatik, zuzenean, 3lb tresnan ireki daitezkeelako,** 3.2.1 atalean azaldutako pauso guztiak saihestuz. Horretarako, siuc01 edo siuc02 makinetan ondorengoa idatziko da:

```
9-[siuc01 PROFIT2]% ./ireki etik1 gezi n
```

Adibide honetan, `ireki` programak zera egingo du:

a. `etik1` karpetan *gezi* hitzaren agerpenak dituen fitxategiak bilatu
b. *gezi* hitza duten fitxategien kopurua eta zerrenda eman
c. eta azkenik, fitxategi hauek bata bestearen atzetik landu nahi diren galdetzen du, aukera ezberdinak ematen dituelarik: **`Bai/Ez/Denak/Atera`**

```
Fitxategi kopurua: 2
      1  eebs.450520392.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)?
```

- **`bai`** idatziz gero –nahikoa da "b" hizkia idatziz gero–, fitxategi hori irekiko da (adibide honetan `eebs.450520392.sat.xml` fitxategia) eta zuzenean etiketatu behar den lema irekiko du (ikus 6. irudia); fitxategi hori etiketatzen bukatzean `File` ➔ `Close` egin eta lana **gorde** ondoren (ikus 8. irudia), programak galdetuko du hurrengo fitxategia ireki nahi den, eta horrela fitxategi guztiekin, bukatu arte:

```
      2  eebs.4715023603.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)?
```

- **ez** idatziz gero –"e" hizkia idatziz gero nahikoa da–, fitxategi hori ez da irekiko eta, zuzenean, hurrengo fitxategia ireki nahi den galdetuko du:

```
Fitxategi kopurua: 2

      1  eebs.450520392.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)? e
      2  eebs.4715023603.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)?
```

- **denak** idatziz gero –"de" idaztea ere nahikoa da–, *gezi* hitza duten fitxategi guztiak irekiko dira 3lb tresnan, hau da, *gezi* hitza duten fitxategi guztiak banan-banan irekiko dira, eta fitxategi bakoitza itxi ondoren, zuzenean, hurrengoa irekiko da, ireki nahi den galderari erantzuten jardun gabe. Aukera honetatik atera nahi bada eta programak ez badu amaitu, `Ctrl Z` sakatu behar da programa eteteko:

```
10-[siuc01 PROFIT2]% ./ireki etik1 gezi n
Fitxategi kopurua: 2

      1  eebs.450520392.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)? denak
Suspended
11-[siuc01 PROFIT2]% jobs
[1]  + Suspended   ./ireki etik1 gezi n
```

Gero idatzi `kill %` eta jarraian `suspended` hitzaren aurretik dagoen zenbakia jarri behar da programa itxi ahal izateko:

```
12-[siuc01 PROFIT2]% kill %1
13-[siuc01 PROFIT2]%
```

- **atera** idatziz gero –"at" soilik idaztearekin nahikoa ere bada–, programa itxiko da.

3.2.2.1. Etiketa edo fitxategi zehatzak irekitzeko

Programa honek bilaketa zehatzagoak ere egin ditzake (synset zehatz batekin etiketatutako hitzak, *Especial Case*[12] zehatz batekin etiketatutako hitzak, fitxategi zehatz bat, eta abar). Nolanahi ere, agindu hauekin **fitxategiak soilik** irekitzen dira. Horregatik, erabilitako etiketaren bat begiratu nahi denean, erabiltzaileak berak bilatu beharko ditu aginduetan zehaztutako etiketa horiek fitxategi bakoitzaren agerpen guztietan.

- Synset zehatz batekin etiketatutako hitzak irekitzeko; horretarako, hitzaren ondoan nahi den synset-zenbakia idatzi behar da:

```
14-[siuc01 PROFIT2]% ./ireki etik1 gezi n 02212344
```

- *Especial Case* zehatz batekin etiketatutako hitzak irekitzeko; horretarako, hitzaren ondoan nahi den *Especial Case*-a idatzi behar da.

[12] *Especial Case*-ak 5.2.1 atalean datoz azalduta.

```
15-[siuc01 PROFIT2]% ./ireki etik1 gezi n C1S
```

- *Especial Case* birekin etiketatutako hitzak irekitzeko; horretarako, hitzaren ondoan nahi diren *Especial Case*-ak idatzi behar dira.

```
16-[siuc01 PROFIT2]% ./ireki etik1 gezi n C3S C4S
```

- Fitxategi zehatz bat irekitzeko:

```
17-[siuc01 PROFIT2]% ./ireki etik1 eebs.450520392.sat.xml
```

Fitxategi osoa idatzi beharrean, fitxategiaren izenaren zati bat bakarrik ere idatz daiteke:

```
18-[siuc01 PROFIT2]% ./ireki etik1 0392
Fitxategi kopurua: 1

      1  eebs.450520392.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)?
```

Aukera honekin kontuz ibili behar da bilaketan idatzi den horixe berarekin (kasu honetan *0392*), fitxategi bat baino gehiagoren izenaren zatia izan daitekeelako.

- *Especial Case* zehatz bat eta synset batekin etiketatutako hitzak irekitzeko; horretarako, hitzaren ondoan lehenengo synset-zenbakia eta ondoren nahi den *Especial Case*-a idatzi behar da.

```
19-[siuc01 PROFIT2]% ./ireki etik1 gezi n 02212344 C1S
```

- Corpuseko hitz baten etiketa jakin batzuk **fitxategi zehatz batetik** aurrera irekitzeko; horretarako, hitzaren ondoan ireki nahi den fitxategi-izena gehituko zaio (berriro ere, fitxategiaren izenaren zati bat bakarrik idatz daiteke):

```
20-[siuc01 PROFIT2]% ./ireki etik1 gezi n eebs.4505
Fitxategi kopurua: 2

      1  eebs.450520392.sat.xml
          Landu nahi duzu (Bai/Ez/Denak/Atera)?
```

Baldintza: fitxategi-izen zatia sartzean gutxienez karaktere alfabetiko bat (zenbakia ez dena) edo punturen bat erabili behar da (bestela programari synset-zenbaki bat irekitzeko agintzen zaio).

3.3. Testuinguruak bilatu `./kwic` *erabilita*

Agindu honek, berez, ez du 3lb interfazea egikaritzen. Baina, 3lb interfazearekin bateragarria da, eta arrazoi honengatik azalduko dugu atal honetan. Agindu honek, 3lb interfazeak ez bezala, **nahi den hitzaren agerpen/testuingurua zuzenean ikusteko** aukera eskaintzen du, hau da, agerpen/testuinguru guztiak jarraian ematen ditu fitxategi izenei erreparatu gabe. Horretarako, siuc01 edo siuc02 makinetan ondorengoa idatziko da:

```
21-[siuc01 PROFIT2]% ./kwic etik1 gezi n
```

Etiketatutako azken hitzak kontuan hartzea nahi bada, `kwic` aginduari `-f` gehitu behar zaio. Bestela azken astean zehar etiketatutakoen berri ez du emango.

```
22-[siuc01 PROFIT2]% ./kwic –f etik1 gezi n
```

Emaitzak hurrengo itxura izango du:

```
23-[siuc01 PROFIT2]% ./kwic -f etik1 gezi n
AGERPEN KOPURUA: 2
ADIEREN MARKAK:
      2 #NOSENSE
ADIEREN AGERPENAK BANAN-BANA:
#NOSENSE: Ikas-norabidean adierazten zaizu, <gezi> baten bidez,
#NOSENSE: Ondoren Alliri zuzendu zizkion <gezi> guztiak.
```

Emaitzan agerpenen kopurua zehazten da, baita agerpenak etiketatutako synset-zenbakia ere. Kasu honetan, *gezi* hitza duten fitxategiak bi dira, eta ez daude etiketatuak, horregatik, `NOSENSE` marka azaltzen da agerpen bakoitzaren aurrean.

```
24-[siuc01 PROFIT2]% ./kwic etik1 begi n
ERABILTZEN DUGUN INDIZEAREN DATA: 2005/2/26 1:32
AGERPEN KOPURUA: 83
ADIEREN MARKAK:
     54 04122028
      5 04122028,C3S
      1 04348269,C3S
     14 C3S
      9 C6S

ADIEREN AGERPENAK BANAN-BANA:
04122028:"Hendaian nago zoraturikan zabal-zabalik <begiak>
04122028: Gaixotasuna krisia da , <begi> zoliagoak.
```

Adibide honetan, ordea, *begi* hitza etiketatua dagoenez synset-zenbakiak ematen ditu (kasu honetan bi esaldi erabili dira adibide gisa, nahiz eta agerpenak guztira 83 diren).

3.3.1. Synset-zenbakiak ikusteko

`./kwic`-ek ematen duen zerrendan agerpen bakoitzaren adiera zein den ikusi nahi izanez gero, aurreko puntuan azaldutako aginduari `-adiera` gehitu behar zaio, honela:

```
25-[siuc01 PROFIT2]% ./kwic -adiera etik1 gezi n
ERABILTZEN DUGUN INDIZEAREN DATA: 2005/3/12 1:31
AGERPEN KOPURUA: 2
ADIEREN MARKAK:
      2 #NOSENSE
ADIEREN AGERPENAK BANAN-BANA:
#NOSENSE:Ikas-norabidean adierazten zaizu , <gezi> baten bidez ,
#NOSENSE:Ondoren Alliri zuzendu zizkion <gezi> guztiak .
```

Agerpen bakoitzaren ezkerretara etiketatzean jarri zaion synset-a adieraziko da adibide honetan ikusi ahal den bezala. Kasu honetan, hitza etiketatu gabe dagoenez `NOSENSE` marka du.

3.3.2. Fitxategi-zenbakiak ikusteko

Honetarako ere beste ikur bat gehitu behar diogu aginduari, ondorengo adibide honetan egiten den bezala:

```
26-[siuc01 PROFIT2]% ./kwic -fitx etik1 gezi n
ERABILTZEN DUGUN INDIZEAREN DATA: 2005/3/12 1:31
AGERPEN KOPURUA: 2
ADIEREN MARKAK:
      2 #NOSENSE
ADIEREN AGERPENAK BANAN-BANA:
eebs.450520392.sat.xml:Ikas-norabidean adierazten zaizu , <gezi> baten
bidez ,
eebs.4715023603.sat.xml:Ondoren Alliri zuzendu zizkion <gezi> guztiak.
```

Agerpen bakoitzaren ezkerretara ikusten den zenbakia, beraz, dagoen fitxategiari dagokiona da.

3.3.3. Ezker eta eskuinera zenbat karaktere nahi diren zehazteko

Zenbaitetan, eskatzen denaren arabera, lerroak nahiko sakabanaturik agertuko dira. Hori ekiditeko, hitzaren ezker eta eskuinera zenbat karaktere agertzea nahi den zehaztu daiteke aginduari `-n` *(karaktere kopurua)* gehituz:

```
27-[siuc01 PROFIT2]% ./kwic -n 50 etik1 gezi n
ERABILTZEN DUGUN INDIZEAREN DATA: 2005/3/12 1:31
AGERPEN KOPURUA: 2
ADIEREN MARKAK:
      2 #NOSENSE
ADIEREN AGERPENAK BANAN-BANA:
               Ikas-norabidean adierazten zaizu , <gezi> baten bidez ,
                    Ondoren Alliri zuzendu zizkion <gezi> guztiak .
28-[siuc01 PROFIT2]% ./kwic -n 10 etik1 gezi n
ERABILTZEN DUGUN INDIZEAREN DATA: 2005/3/12 1:31
AGERPEN KOPURUA: 2
ADIEREN MARKAK:
      2 #NOSENSE
ADIEREN AGERPENAK BANAN-BANA:
n zaizu , <gezi> baten bid
u zizkion <gezi> guztiak .
```

Ikusten denez, lehenengoan hitzaren ezker-eskuinera 50 karaktere agertzea eskatu da, bigarrenenean aldiz, 10. Informazio ugari eskatu nahi denean baliagarria izan daiteke hau, guztia argiago ikuste aldera.

3.3.4. Aginduen hurrenkera

a. Lehenengo: `./kwic`
b. Ondoren, indizea berritzeko agindua, `-f,` aukeran.
c. Hurrengo hauei nahi den hurrenkera eman dakieke, eta hautazkoak dira: `-n, -adiera, -fitx.`
d. Taldeko kide bakoitzari dagokiona: `etik1/etik2/epai`

e. Hitza
f. Kategoria: `n`, `a` edo `v`

Lehenengoa, eta azken hirurak ezinbestean idatzi behar dira. Adibidez:

```
29-[siuc01 PROFIT2]% ./kwic –f -adiera -fitx -n 50  etik1 gezi n
LEHENBIZI INDIZEA BERRITZEN...
AGERPEN KOPURUA: 2
ADIEREN MARKAK:
      2 #NOSENSE
ADIEREN AGERPENAK BANAN-BANA:
eebs.450520392.sat.xml:  #NOSENSE:Ikas-norabidean  adierazten  zaizu  ,
<gezi> baten bidez,
eebs.4715023603.sat.xml:   #NOSENSE:Ondoren   Alliri   zuzendu   zizkion
<gezi> guztiak .
```

3.3.5. Emaitzak *emacs* edo *xemacs* editorean egikaritzeko

Esan bezala, lerroak luze samarrak gerta litezke, eta `./kwic` aginduari eman nahi zaion erabileraren arabera, desitxurosoak. Hau honela izan ez dadin, aginduak *emacs* edo *xemacs* editorean egikaritu beharko dira.

Honetarako, lehenik eta behin, editorea zabalduko dugu *shell*-etik bertatik. Bi editore hauek dira baliagarriak:

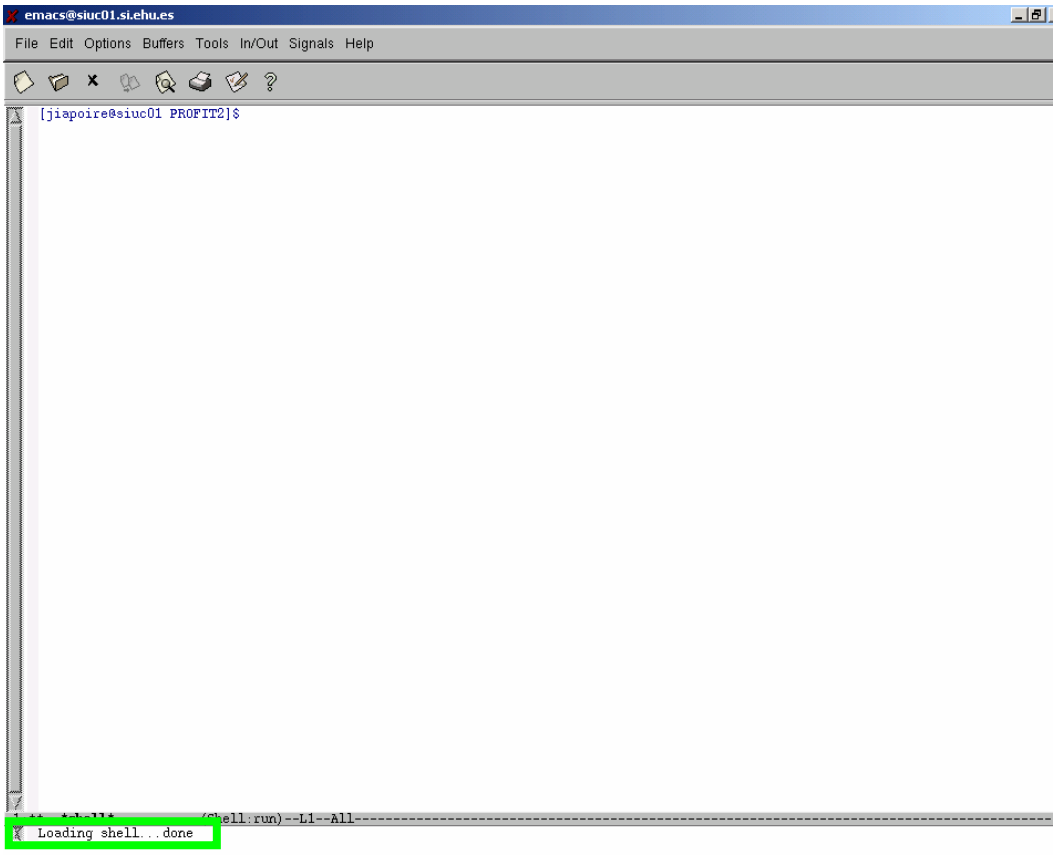
```
30-[siuc01 PROFIT2]% emacs &
```

edo

```
31-[siuc01 PROFIT2]% xemacs &
```

Editorearen leihoa zabaldu eta hurrengo pausoa *Esc* eta *x,* biak batera sakatzea izango da; beheko lerroan idazteko aukera emango digu honek. Bertan, *shell* idatziz gero, komandoak exekutatzeko leihoa zabalduko zaigu, hau da, *shell*-a (ikus 9. irudia)[13].

[13] Bai *emacs* eta bai *x-emacs* erabili daitezkeen arren, txosten honetako adibide guztiak *emacs* editorearenak izatea erabaki dugu.



9. irudia

Leiho honetan lerroak oraindik luzeegiak dira, eta hauek hobeto ikusteko berriro ere *esc* eta *x* sakatu. Honako aginduak idatzi behar dira gero:

- `set-variable <return>`
- `truncate-lines <return>`
- `t <return>`

3.3.6. Emaitza inprimatzeko

Lerro-zerrenda inprimatu nahi bada, `./kwic` agindua modu honetan idatzi:

```
32-[siuc01 PROFIT2]% ./kwic -fitx etik1 gezi n | lpr -P lsi_2p
```

Hau da, `| lpr -P lsi_2p` gehitu behar zaio hasierako aginduari.

****

Oharra: Orain arteko adibide guztietan `etik1` erabili den arren, kontuan hartu behar da etiketatu behar duten beste kideek berdin-berdin erabiliko dituztela programa hauek, baina `etik1` jartzen duen lekuan `etik2` (2. etiketatzaileari dagokiona) edo `epai` (epaileari dagokiona) idatzi beharko dute.

4. Editatze-lana

Editorea EuskalWordNet ezagutza-basean *editatzen* duen pertsona da, hots, ingeleseko WordNet-eko[14] synset-etan oinarrituaz, euskarako WordNet-ean kontzeptuak txertatzeaz/orrazteaz arduratzen dena. Ondoren, editorearen eginbeharrak zehaztuko dira[15]. Eginbeharren azalpena ulerterrazagoa egiteko, eta asteko bilera ardatz gisa hartuta[16], eginbeharrak bi fasetan banatuak izan dira:

a. Asteko bilera baino lehenagoko eginbeharrak
b. Asteko bileraren ondorengo eginbeharrak

4.1. Asteko bilera baino lehenagoko eginbeharrak

Etiketatzaileei eta epaileari asteroko bileran hitzak banatu baino lehen, editoreak hitz hauek landu egin behar ditu, ondorengo pausuak jarraituz:

1. Astero, zerrendatik hainbat hitz aukeratuko ditu: 2.2 atalean aipatu bezala, hitz-zerrenda bitan banatua dago (`landutakoak.n.usuenak.txt` eta `landutakoak.n.ezusuak.txt`). Baina, egun, usuen zerrenda bakarrik (`landutakoak.n.usuenak.txt`) baliatzen da, hau bukatu arte behintzat.

Fitxategi hauetara iristeko *SSH Secure Shell Client*-en[17] ondorengoan idatzi behar da:

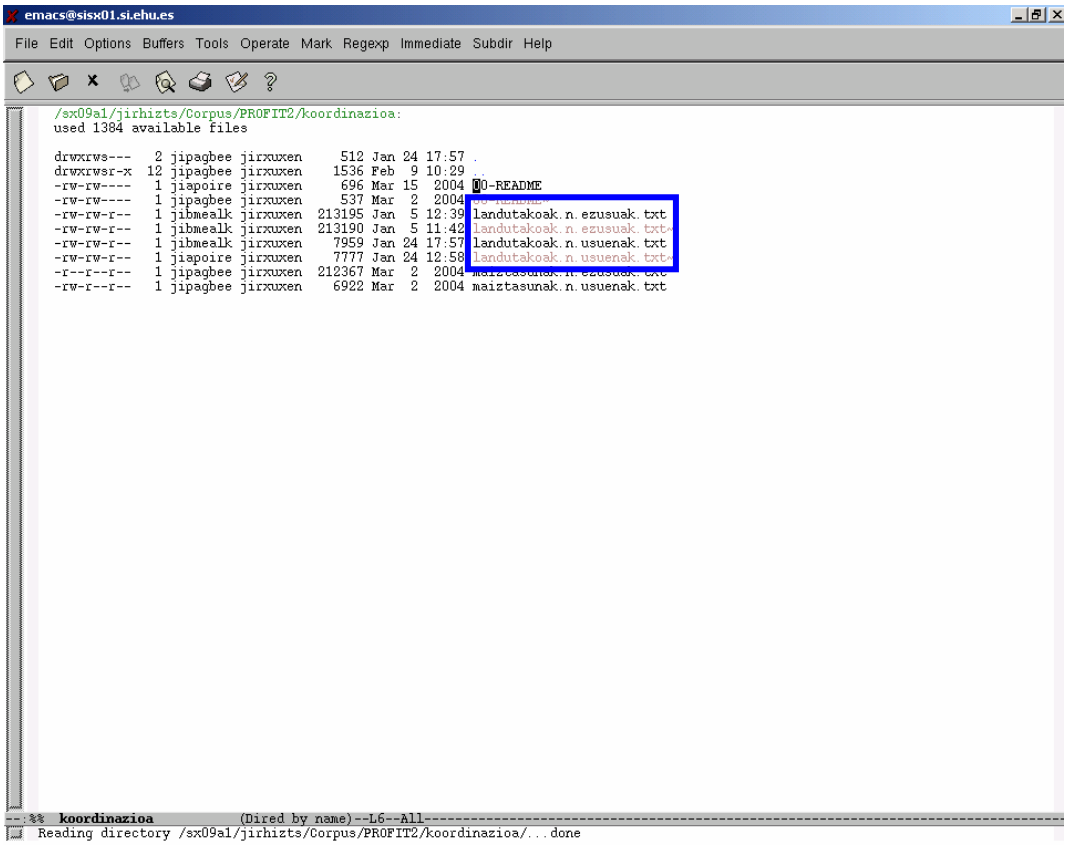
```
33-[siuc01]% cd ~jirhizts/Corpus/PROFIT2/koordinazioa
```

Ondoren, *emacs* (edo *xemacs*) interfazea ireki behar da. Horretarako, idatzi:

```
34-[siuc01 PROFIT2]% emacs &
```

Emacs irterfazean `File` ➔ `Open Directory` eta ondoren *return* sakatuta, bertan egongo dira bi hitz-zerrenden fitxategiak ikusgarri (ikus 10. irudia):

[14] Normalean, ingeleseko WordNet-a hartzen da ardatz gisa, baina gaztelaniako WordNet-a ere erabili ohi da.
[15] Kontuan izan beharrekoa da, eskuliburu honetan aipatzen diren editorearen eginbeharrak, EuSemCor corpusa semantikoki etiketatzeko proiektuaren barnean kokatzen direla. Baina, editore lana guztiz osatzeko, editoreak *EuskalWordNet-en orrazketa: Editorearen eskuliburua* ere baliatu beharko du (txosten honen A eranskinean).
[16] Ikus 2.3 atala bileren berri izateko.
[17] *Shell*-a egikaritzeko ikus 3. atala.



10. irudia

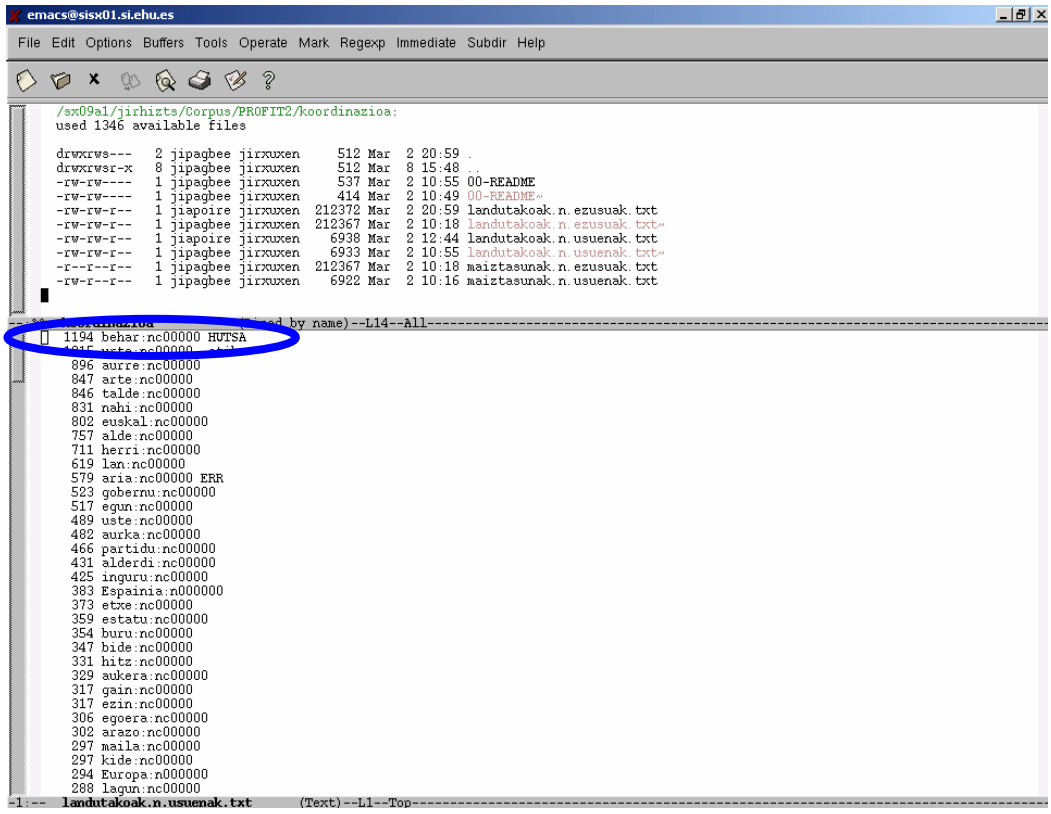
Ireki nahi den fitxategiaren gainean kokatu[18], eta *return* sakatu fitxategira sartzeko. Azkenik, zerrenda horretatik hitz batzuk aukeratzea[19] bakarrik falta da. Hala ere, zenbait hitz *salbuespen* gisa tratatzen dira; esate baterako, hitzen bat **EuskalWordNet eta ingeles/gaztelaniako WordNet-etan ez badago**, **zenbakia, izen berezia, edo postposizio/perifrasi baten osagaia** bada, alde batera utziko da, baita **gaizki lematizatua** dagoenean ere. Horrela bada, editoreak hitz hauei marka bat ezarriko die bere ondoan:

- **ERR** (lematizazio errorea)
- **IZB** (EuskalWordNet-en eta ingeles/gaztelaniako WordNet-etan ez dagoen izen berezia, oraingoz ez landu)

[18] Gorriz dauden fitxategiak, aldi baterako fitxategiak dira, horiek ez dira aukeratu behar.

[19] Normalean, zerrendako hitzak agerpenen ordena jarraituta aukeratuko ditu. Hala ere, une honetan, IXA taldeko kide batzuk `landutakoak.n.ezusuak.txt` zerrendatik EuskalWordNet-en hitz monosemiko gisa agertzen direnak aztertzen ari dira, hauek benetako monosemikoak diren ala ez egiaztatzeko (honen berri 7. atalean emango da). Lan honen ondorioz, marka berriak sortu dira: **`AUTO`** eta **`AUTOC1S`**. Lehenengo marka daramaten hitzak monosemikoak dira eta automatikoki etiketatuko dira; bigarren marka daramatenak, aldiz, monosemikoak ez diren hitz horiei ezarriko zaie –polisemikoei, alegia–, eta editoreak landu beharko ditu, gero, etiketatzaileek hauek eskuz etiketatu ahal izateko. Hortaz, EuskalWordNet-en hitz monosemiko gisa agertzen direnak aztertzen amaitzean, etiketatzeko dauden hitz askok, dagokien zerrendan, `AUTO` edo `AUTOC1S` marka eramango dute ondoan, eta editoreak `AUTOC1S` marka daramaten hitzak aukeratu beharko ditu zuzenean, hau da, monosemikoak ez izaki, automatikoki etiketatu ezin diren horiek, alegia. Bi marka hauei buruz, 7. atalean mintzatuko gara luzeago.

- **HUTSA** (postposizio edo aditz perifrastikoak: *ahal, behar, aurre, alde...*)
- **EWN** (hitza ez dago EuskalWordNet-en, ezta ingeles/gaztelaniako WordNet-etan ere : *sagardotegi, euskal...*)[20]
- **ZKI** (digituak)



11. irudia

11. irudian ikus daitekeen bezala, fitxategian bertan editoreak *behar* hitza HUTSA bezala (postposizio edo aditz perifrastiko bezala, alegia) markatu du.

2. Editoreak etiketatzaileek etiketatuko dituzten hitzak aukeratu ondoren, hitz horiek EuskalWordNet-en dituzten synset-ak aztertu/orraztuko ditu eta ez daudenak gehituko ditu, hurrengo baliabideak erabilita:
    a. EuskalWordNet 1.6 (http://siuc02.si.ehu.es/cgi-bin/mcrWei/privat/wei.edit.perl)
    b. *Hiztegixa* (http://ixa2.si.ehu.es/hiztegixa)
    c. EDBL Datu-base lexikala: (http://ixa2.si.ehu.es/edbl/)
    d. Euskarako hiztegiak:
        i. *Euskalterm* (http://www1.euskadi.net/euskalterm)
        ii. *Elhuyar Hiztegi Txikia*
        iii. *Elhuyar Hiztegia (elebiduna)*
        iv. *Euskal Hiztegi Modernoa*
        v. *Euskal Hiztegia*
        vi. *Euskaltzaindia* (http://www.erabili.com/lantresnak/hiztegiak/euskaltzaindia)
        vii. etab.

[20] Hitz bat ez badago WordNet-etan, zerrenda batean apuntatu behar da (argibide gehiago 4.2 eta A.3.2.7.2 ataletan)

e. Ingeleseko hiztegiak:
    i. *Oxford* (paperean)
    ii. *Collins* (paperean)
    iii. *Wordreference* (http://wordreference.com)
    iv. *Cambridge* (http://dictionary.cambridge.org)
    v. *Onelook* (www.onelook.com)
    vi. etab.
f. Corpusak:
    i. *XX. mendeko Euskararen Corpus Estatistikoa* (http://www.euskaracorpusa.net/XXmendea/)
    ii. *Ereduzko prosa gaur* (http://www.erabili.com/lantresnak/aztergailuak/prosa)
    iii. *EuSemCor corpusa* (3lb tresna edota `./kwic` agindua erabilita)[21]
    iv. *Google*
g. EuskalWordNet orrazteko editorearen eskuliburua

3. Behin hitz baten synset-ak landuta, hauek inprimatzen dira, etiketatzaileei eta epaileari banatu beharreko fitxak prestatzeko. Fitxa hauetan synset-ez gain, honako informazioa txertatu behar du editoreak:
    a. Synset-en hiperonimo eta hiponimoak, synset-ek adierazten dutena hobeto ulertu ahal izateko.
    b. Synset horiei adibide/testuinguruak erantsi. Horretarako, ondorengo baliabideak erabiliko ditu:
        i. Ingeleseko WordNet 1.6. Hau exekutatzeko:
            1. `[sisx01]% wn16`
            2. `[sisx01]% wnb &`
        ii. Ingeleseko WordNet 2.0 (http://wordnet.princeton.edu/cgi-bin/webwn)
        iii. Corpusak:
            1. *XX. mendeko Euskararen Corpus Estatistikoa*
            2. *Ereduzko prosa gaur*
            3. *EuSemCor corpusa*
            4. *Google*
    c. *Elhuyar Hiztegi*ko hitz horren adierak eta HAULak.
    d. Beharrezkoa iruditzen zaionean, adiera oso antzekoa duten synset-ak elkartzeko proposamena adierazita eraman[22].

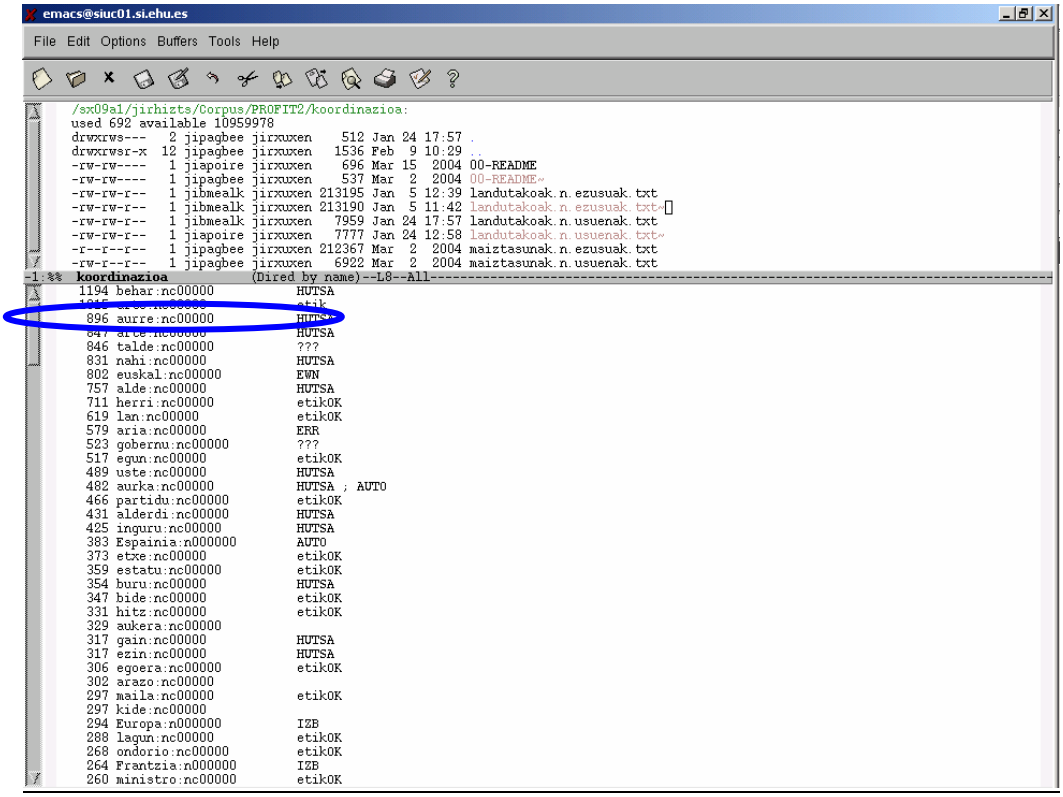
[21] Ikus 3. atala.
[22] EuSemCor etiketatzean, antzeko adiera duten synset-ak bateratu daitezke, hau da, agerpen bati bi etiketa ematea badago, hauek testuinguruan bereiz ezinak baitira (argibide gehiago A.3.2.6.4 atalean).

4.2. Asteko bileraren ondorengo eginbeharrak

1. Asteko bileran bertan erabakitzen diren EuskalWordNet-eko aldaketa guztiak egingo ditu:

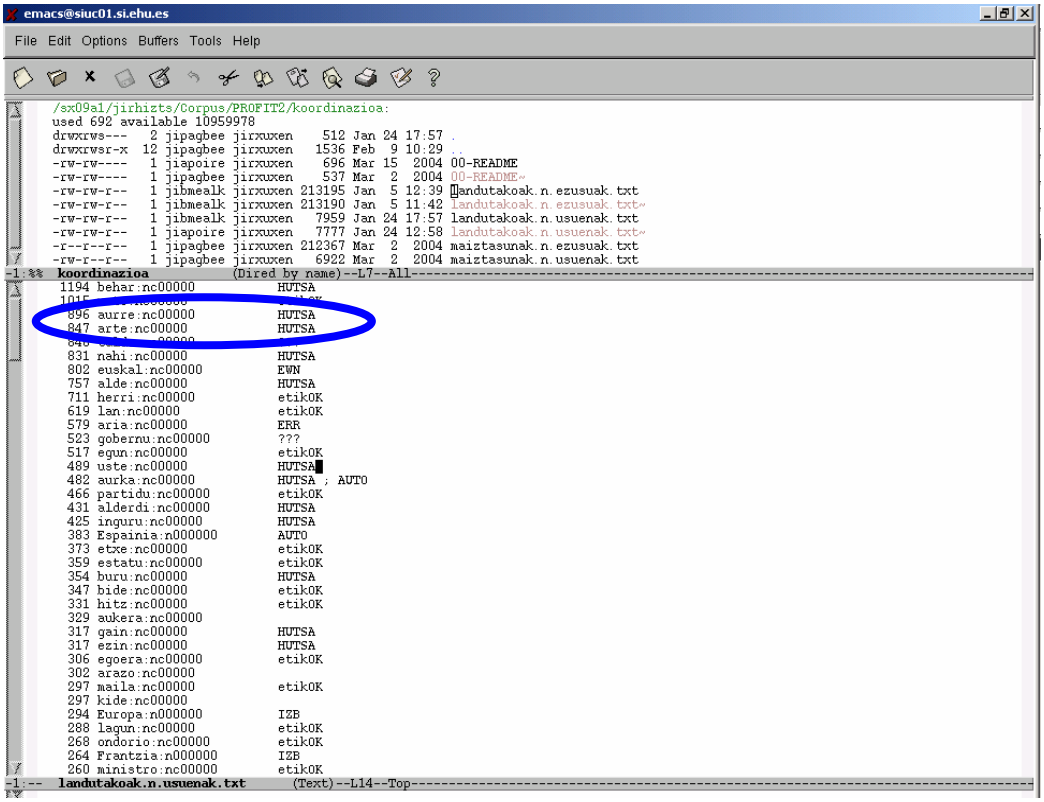
    a. Etiketatzen hasi baino lehen, bileran bertan adieraren bat falta dela ohartuz gero, editorea adiera hori EuskalWordNet-en sartzen saiatuko da.
    b. Hasierako synset zerrendatik synset-en bat kentzen bada, editoreak EuskalWordNet-etik ere kenduko du.
    c. Baliabideak: (ikus 4.1 ataleko 2. eta 3. puntuak)

2. Editoreak etiketatzaileei eta epaileari etiketatu beharreko hitzak **banatu ondoren, `etik` marka emango die dagokien zerrendan** (ikus 4.1 atala). 12. irudian ikus dezakegun bezala, editoreak *urte* hitzari `etik` marka jarri dio, eta honek **hitza etiketatzeko banatua izan dela** adierazten du. Editoreak `etik` marka aldatuko du, eta **`etikOK`** marka jarri, etiketatzaileek hitz bakoitza etiketatzen amaitu dutela jakinarazi bezain laster. Horrela, epaileak `etikOK`[23] marka ikusten duenean, badaki hitz hori prest dagoela epaitzeko. 13. irudian ikus daitekeen bezala, *urte* hitzaren `etik` marka orain `etikOK` da, horrela, hitz horren etiketatze-lana zeharo amaitu dela, eta epaileak epaitu dezakeela jakin daiteke.



12. irudia

[23] C eranskinean hitzei ematen zaizkien marka guztiak zerrendatuak datoz.



13. irudia

3. Bileran epaileak corpusean agertu diren adiera berriak jakinarazi dizkio editoreari[24], eta honek EuskalWordNet-en sartuko diren ala ez erabakiko du[25]. Gainera, editoreak bere erabakiaren berri eman beharko dio epaileari[26]. Zalantzaren bat izanez gero, lan-taldeari eskatuko dio laguntza. WordNet ezberdinek (ingelesekoak, gaztelaniakoak eta euskarakoak) synset hori izango ez balute, editoreak hitz horren adiera jakin hori falta dela apuntatuko du *EuskalWordNet-en ez dauden adierak* zerrendan edota *EuskalWordNet-en ez dauden hitzak* zerrendan –informazio gehiago EuskalWordNet orrazteko eskuliburuan (ikus A.3.2.7.2 eta A.3.2.7.3 atalak A eranskinean). Adiera berriak aztertzean, *polisemia erregularra* kontuan hartu behar da, ikus 4.2.1 atala.

4. Bileran epaileak corpusean agertu diren HAULak jakinarazi dizkio editoreari, epaileak agerpen horiek, argi eta garbi, HAULak direla onartu ondoren (ikus 6.4). Hauek EuskalWordNet-eko synset batekin lotzen saiatuko da, eta lan-taldeari bere erabakiaren berri emango dio. WordNet ezberdinek HAUL batentzako synset egokirik izango ez balute, editoreak EuskalWordNet-en HAUL jakin hori falta dela apuntatuko du *EuskalWordNet-en ez dauden HAULak* zerrendan.

[24] Adiera berria behar duten hitzek, **`epaiOKC1S`** marka eramango dute `landutakoak.n.usuenak.txt` fitxategian. Marka hau epaileak jartzen du (ikus 6.4 atala).
[25] Horretarako hiztegiak eta corpusak baliatuko ditu.
[26] Adiera berri hori EuskalWordNet-en txertatuz gero, epaileak jakin behar du, corpusean adiera horrekin agertu diren agerpenak etiketatu behar dituelako (ikus 6.4 atala).

5. Arrazoi batengatik, editore, epaile eta etiketatzaileek erabakiko balute hitz bat geroago etiketatuko dutela, editoreak hitz horri **???** marka gehituko dio dagokion fitxategian (hori egiteko ikus 4.1 atala). Adibidez, 13. irudiko *talde* hitza, geroago etiketatzeko utzi da.

6. Arrazoi batengatik, editore, epaile eta etiketatzaileek erabakiko balute hitz baten synset bat ez dela egokia, editoreak ezingo du hitz hori synset-zerrendatik kendu epailearen baimena izan arte[27].

7. Epaileak epaitu beharreko hitza adostasun maila baxukoa bada, asteko bilerara eramango du, eta hango erabakien arabera editoreak hitz hori berriro landu beharko du.

8. Hitz bakoitzarekin igarotako denbora apuntatuko du (bilera aurrekoa + bilera ondorengokoa).

4.2.1. Polisemia erregularra

Lexiko semantikan *polisemia erregularra* deritzon fenomenoa oso arrunta da, hitz batzuek jasaten duten adiera-alternantzia erregularra da, eta gainera, aurreikus daitekeena. Kategorien artean, adjektiboak dira fenomeno honen adibide garbiena, hauek ondoan hartzen duten izenen arabera esanahi oso desberdinak izan baititzakete. Esate baterako, ingeleseko *fast* adjektiboak, hurrengo adierak izan ditzake :

> The adjective fast in receives different interpretations when modifying the nouns *programmer*, *plane* and *scientist*. A *fast programmer* is typically a programmer who programs quickly, a *fast plane* is typically a plane that flies quickly, a *fast scientist* can be a scientist who publishes papers quickly, who performs experiments quickly, who observes something quickly, who reasons, thinks, or runs quickly. Interestingly, adjectives like *fast* are ambiguous across and within the nouns they modify. A *fast plane* is not only a plane that flies quickly, but also a plane that lands, takes off, turns, or travels quickly. Even the more restrictive *fast programmer* allows more than one interpretation. One can easily think of a context where a fast programmer thinks, runs or talks quickly.
>
> (Lapata, 2001)

Hala ere, fenomeno hau ez da adjektiboekin bakarrik gertatzen, izenetan ere ikus baitaiteke. Adibidez, herrien izenek hainbat adiera izan ditzakete:

a. *Holanda polita da*.
b. *Holandak irabazi zuen*.
c. *Holandak aukeratu du ordezkari hori*.

Lehenengo esaldian, *Holanda* estatu-izena adierazi nahi da; bigarrenean, *Holanda* herriaren ordezkaria den pertsona-multzo bat (kirol-talde bat, adibidez); eta azkeneko esaldian, holandarrei buruz ari da.

[27] 6. atalean azalduko dira erabaki honen arrazoiak.

Nahiz eta oso fenomeno arrunta izan, hiztegiek ezin dituzte adiera hauek guztiak jaso, hauek testuinguruaren baitan baitaude. EuskalWordNet-en ere horrelako hutsuneak arruntak dira, eta EuSemCor etiketatzean, maiz agertu izan dira horrelako adibideak. Hala ere, proiektu honen lehenengo fase honetan, polisemia erregularra alde batera utziko da, hurrengo urrats batean lantzeko asmoarekin.

Hala, editoreak adiera berrien artean horrelako kasuak baditu[28], beste hizkuntzetako WordNet-etan dauden begiratuko du:

a. Adiera hori jasotzen duen synset-en bat badago, bertan txertatuko du adiera berri hori.

b. Ez badu adiera hori jasotzen duen synset-ik topatzen:

- EuskalWordNet-en ez dauden adierak zerrendan apuntatuko du (ikus A.3.2.7.2 atala), eta ohartxo bat gehituko dio polisemia erregularra dela esanez, **eta gainera, corpusean agerpen horiei emango zaien synset-a ere apuntatuko du.**

- **Epaileari emango dio erabaki horren berri, eta epaileak etiketatuko ditu corpuseko agerpen horiek.**

[28] Etiketatzaileek corpusean adiera berri gisa ikusten duten guztia marka dezaketen arren, azkeneko erabakia editoreak hartuko du.

5. Etiketatze-lana

Etiketatzailea EuSemCor corpusa *etiketatzen* duen pertsona da. Hala ere, etiketatzaileak ez ditu corpuseko hitz guztiak etiketatuko, etiketatze-prozesua erdiautomatikoa baita, hau da, monosemikoak diren hitz guztiak, hurrengo fasean, automatikoki etiketatuko dira (ikus 7. atala). Hortaz, etiketatzaileak eskuz etiketatuko ditu bakarrik hitz polisemikoak[29].

Gainera, prozesuaren metodologia hitz polisemikoen agerpen guztiak jarraian etiketatzean oinarrituko da, hau da, etiketatzerakoan ez dira fitxategietako esaldi guztiak jarraian aztertuko; aitzitik, aztergai den hitz polisemikoa duten esaldiak/fitxategiak bakarrik etiketatuko dira.

Ondoren, etiketatzailearen eginbeharrak zehaztuko dira. Azalpen hau ulerterrazagoa egitearren, eginbeharrak hiru fasetan banatuak izan dira:

a. Etiketatu baino lehenagoko eginbeharrak
b. Etiketatzean egin beharrekoak
c. Etiketatu ondorengo eginbeharrak

5.1. Etiketatu baino lehenagoko eginbeharrak

1. Etiketatzailea, epailea eta editorearekin asteroko bileran elkartuko dira eta editoreak etiketatzaileari etiketatu beharreko hitzak banatuko dizkio (ikus 2.3 atala). Bilera horretan etiketatzailea editoreak planteatzen dituen synset-ak ulertzen/adosten saiatuko da.

2. Etiketatzaileak, hitz baten agerpenak etiketatzen hasi baino lehen, editoreak banatutako hitz horren *fitxa* errepasatuko du.

5.2. Etiketatzean egin beharrekoak

1. Behin hitzaren synset-ak ulertuak/adostuak, etiketatzaileak hitz horren agerpenak etiketatuko ditu, eta horretarako, 3lb tresna erabiliko du (ikus 3. atala).

2. Etiketatzaileak hitzak corpusean duen agerpen bakoitzari, gutxienez, synset bat edo *Especial Case* bat egokituko dio.

3. Etiketatzen ari den bitartean sortu zaizkion zalantza pertinente guztiak apuntatuko ditu, lan-taldearekin egingo den hurrengo bileran aztertzeko.

[29] Hitzak monosemikoak/polisemikoak diren, EuskalWordNet-en dituzten synset-en arabera erabakitzen da. Hala ere, 7. atalean azalduko den bezala, honek ez du ziurtatzen hitz horrek benetan adiera bakarra duenik.

4. Etiketatu ahala, etiketatzaileak synset bakoitzerako adibide adierazgarrienak apuntatuko ditu, gerora, itzulpenetan gehitu ahal izateko[30].

5.2.1. Kasu bereziak edo *Especial Case*-ak

Especial Case deitzen zaie editoreak EuskalWordNet-etik proposatu dituen synset horiekin etiketatu ezin diren agerpenei. Hala, etiketatzaileak editoreak proposaturiko synset horiek etiketatzeko ezin dituenean baliatu, *Especial Case* markak erabiliko ditu. Zazpi *Especial Case* ezberdin daude:

5.2.1.1. Especial Case 1: Word exists in dictionary[31] but not its sense

Adiera berriei ezartzen zaien marka da, hau da, etiketatzaileak agerpen bateko adiera editoreak emandako synset-zerrendan ez dagoela uste badu, agerpen hau *Especial Case 1* bezala markatuko du.

Ikus dezagun adibide bat. Demagun etiketatu beharreko hitza *artikulu* dela, eta editoreak emandako hitz honen synset-zerrenda hurrengoa dela:

communication LanguageRepresentation Part Tops	3 **article_4** lock 0 **artikulu_4** lock 2 **artículo_5**	a determiner that may indicate the specificity of reference of a noun phrase kasua eta numeroa erakusten duen eta izen zein adjetiboari itsasten zaion determinatzaile mota Clase de determinante que indica el género y el número del nombre que acompaña
communication Agentive Communication Manner Mental Purpose Social Tops UnboundedEvent	11 **article_1** lock 5 **artikulu_7** lock 8 **artículo_4**	nonfictional prose forming an independent part of a publication argitalpen bat osatzen duen prosazko zati independiente bat; "artikulu#errezetak ematea ez da artikulu honen helburua, aldiz, informatikak hizkuntzen ikaskuntza arloari eskaintzen dizkion posibilitateen ikuspegi orokorra izatea da" Escrito que aparece junto con otros en una publicación

14. irudia

Etiketatzaileak synset hauetaz gain, corpusean beste adiera berriren bat aurki dezake, esate baterako, "4. artikuluak ezartzen duen printzipioaren argitan: Askatasun zibilaren oinarriaren arabera..." esaldia. Kasu honetan, testua legeari buruz ari da, eta synset-zerrendan ez dagoenez adiera hori, etiketatzaileak *Especial Case 1* marka erabili beharko du. Gainera, etiketatzaileak asteko bileran adiera berri horien berri eman ahal izateko, hauek agertu ahala, apuntatu beharko ditu.

[30] 5.3 atalean ikusiko dugun bezala, etiketatze-lana amaitzean etiketatzaileak synset-en glosak itzuli egiten ditu.
[31] *Dictionary* esatean EuskalWordNet (edo WordNet) adierazi nahi da.

5.2.1.2. Especial Case 2: Word does not exist in dictionary

Marka hau etiketatu beharreko **hitza** EuskalWordNet-en ez dagoenean erabiltzeko sortu zen. Hala ere, *Especial Case* hau **ez da inoiz gertatuko**, editoreak epaile eta etiketatzaileei emandako hitz guztiak EuskalWordNet egongo direlako.

5.2.1.3. Especial Case 3: Word is part of a Multiword Lexical Unit or is a lexicalized inflected form

Kasu berezi hau hitz-anitzeko unitate-lexikalak (aurrerantzean, HAUL) markatzeko erabiltzen da. Gerta daiteke etiketatu beharreko hitza, agerpenen batean, HAUL baten osagaia izatea. Esate baterako, *adarra* hitza etiketatzean, etiketatzaileak corpusean "Gobernuari adarra jo nahiean..." testuingurua etiketatu beharko balu, *Especial Case 3* marka erabiliko luke. Horrela, *adarra* hitza HAUL baten (*adarra jo*ren) osagaia dela adierazten da, hots, agerpen horretan *adarra* hitza ondoko hitzarekin batera unitate bat osatzen duela[32].

Hona hemen HAULen adibide batzuk: *urte berri, polizi judizial, alderi politiko, boto-txartel*[33]*, hitz egin, hitza eman, kultur etxe, euskal etxe, gizon-dantza, kale egin, kale itsu, kalera bota, neska-lagun, mutil-lagun, zine-zuzendari, orkestra-zuzendari, ...*

Etiketatzaileak HAULa izan daitekeen agerpen bat etiketatu behar duenean, ondorengo pausoak jarraitu behar ditu:

a. Agerpen hori *Especial Case 3* bezala markatuko du.

b. **Ahal baldin bada**, agerpenari synset bat ematen saiatuko da[34]. Hala ere, kasu batzuetan HAULaren osagai bakoitzaren adierak banatzea ezinezkoa da: *adarra jo*n, adibidez, zer esan nahi du *adarra*k kasu horretan? Zaila da definitzen, HAULa bere osotasunean hartzen ez bada. Horregatik, etiketatu beharreko agerpenari (*adarra*, adibidez) **synset bat bera ere ez badagokio**, ez zaio jarri behar *Especial Case 1* marka, hau da, ***Especial Case 3* eta *Especial Case 1* markak ez dira bateragarriak**[35].

---

[32] 3lb tresnan gorriz markatuta dauden hitzak –etiketatu beharreko hitzak, alegia– lematizatzailetik (EDBLtik) datoz. Gaurko EDBL bertsioarekin, lematizatzaileak ezin ditu HAULak bereizi, horregatik sortu behar izan dugu *Especial Case* hau.

[33] Hemen ohar bat egin beharra dago. EDBLn marratxodun hitz konposatuak ez dira HAULtzat hartzen, bai, ordea, haien pareko marratxorik gabeak (*boto txartel*). Guk hemen ez dugu bereizketa hori egingo eta biak izango dira HAUL (aurrerago, HAULekin zer egin erabakitzen dugunean, gogoan izan beharreko kontua da).

[34] Irizpide hau era berean beteko da hitz-elkarketan parte hartzen duten *udal-, herri-, mendi-* eta abar bezalako izenekin. Horrelakoak ingelesera edo gaztelaniara itzultzean, izena ez den beste kategoria bat (adjektiboa) hartzen dutelako (***herri**-eskola = escuela **pública**; **udal**-barrutia = distrito **municipal**; **mendi**-aterpe = refugio **de montaña***). Beraz, euskarako izen hauek, WordNet-eko adjektibo synset batzuei dagokie. A.3.2.7.1 atalean ikusiko dugun bezala, oraingoz alde batera utziko ditugu adjektiboak landu arte. Baina, etiketatzaileek adjektibo gisa itzultzen diren eta beti hitz-elkarketan parte hartzen duten izen hauek, beste HAULak bezala (*Especial Case 3* eta synset batekin) etiketatuko dituzte.

[35] Etiketatu beharreko agerpenari, *Especial Case 3* eta *Especial Case 1* markak batera ezartzeko beharra askotan gertatzen denean, hitz horren adiera *usu* bat falta denaren adierazgarri da. Horrelakoetan, hitz horren etiketatzeari utzi, eta editoreari lehenbailehen jakinaraziko zaio.

Salbuespena: *Especial Case 3* eta *Especial Case1* markak bateragarriak dira hurrengo kasuan. Zenbaitetan izen baten adiera *usu* hori ez dago WordNet-en (normalean, euskarazko kontzeptu kulturalak

c. Eman dion synset-a apuntatuko du (ez badio synset-ik eman ere), gero ager daitezkeen HAUL horren beste agerpenetan synset berarekin markatzeko. **Horrela, HAUL bera beti berdin etiketatuko du**.
d. Gerta liteke, HAUL batek adierazten duen adiera bera, HAULeko osagai **bakar batek** adieraztea, elipsiaren antzeko zerbait gertatzea, alegia:

- Partidu politiko guztiek uka dezatela...
- Partidu guztiek uka dezatela...

Etiketatzaileak *partidu* hitza etiketatzen ari bada, lehenengo kasuan (*partidu politiko* agertzean, alegia) a., b. eta c. pausoak jarraituko ditu, hau da, agerpen hori *Especial Case 3* bezala etiketatuko du, synset bat ematen saiatuko da eta emandako synset-a apuntatuko du[36]. Bigarren adibidean (*partidu* bakarrik agertzean), bi aukera egon daitezke:

a. Editoreak banatutako synset-zerrendan *partidu politiko* eta *partidu*, **biak** synset berean badaude (ikus 15. irudia), orduan, etiketatzaileak, zuzenean, synset horrekin etiketatuko du agerpen hori (kasu honetan, *partidu_2*-rekin).

base concept		an organization to gain political power
group	29 **party_1 political_party_1**	botere politikoa erdiestea helburu duen erakundea;
Function	lock 0 **partidu_2 partidu_politiko_1**	"partidu#1992an, nazio-mailan hirugarren partidu bat
Group	**alderdi_politiko_1 alderdi_2**	antolatzen saiatu zen Perot"
Human	lock 42 **partido_2 partido_político_1**	Organización política cuyos miembros comparten la misma ideologí

15. irudia

b. Editoreak banatutako synset-zerrendan *partidu politiko* eta *partidu* biak synset berean egongo ez balira (ikus 16. irudia), etiketatzaileak *ezkutuan* HAUL bat dagoela adierazteko, a., b. eta c. pausoak jarraituko ditu, hau da, agerpen hori

---

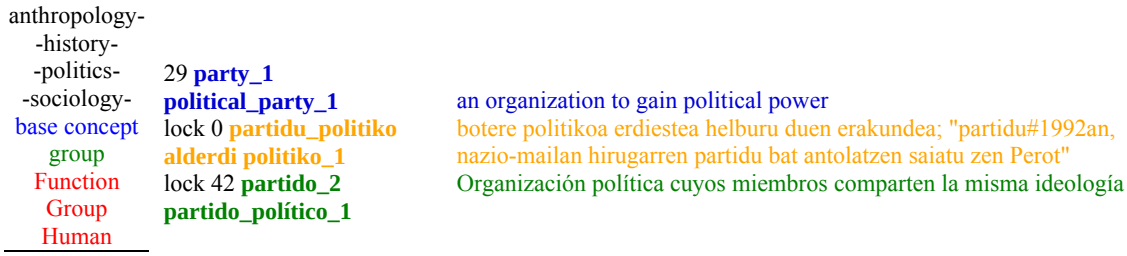
dira: *pilota, ikastola, kalimotxo…*). Hau da, euskarazko izen bat polisemikoa izan daiteke, baina izen horren adiera arrunt (*usu*) bat euskal kontzeptu bati dagokio, eta honek ez du synset-ik WordNet-en. Ikus dezagun adibibe bat: *pilota* izena *Euskal Hiztegi Moderno*an hurrengo adierekin definitzen dute:

> iz. **A1** "Material elastikozko objektu biribila, gogorra zein biguna, barrutik hutsa zein betea eta diametro ez oso handikoa dena (hainbat kiroletan erabiltzen da, horien artean euskal pilota-jokoa dagoelarik)" *Tenis-pilota*
> iz. **A2** "Delako objektuaz jokatzen den joko edo kirola, jatorriz Euskal Herrikoa eta hainbat modalitate desberdin dituena"
> iz. **A3** "Antzeko ezaugarriak dituen objektu edo gauza" *Elur-pilota. Gomazko pilotak.*

Hiru adiera hauetatik bik (A1 eta A3) synset-a dute EuskalWordNet-en. Aldiz, "jokoa edo kirola" dagokion adieraeri (eta corpusean gehien agertzen denari) ezin izan zaio synset bat egokitu, WordNet-en ez baitago hori adierazten duen kontzepturik. Horrelakoak agertzean, 5.2.1.1 atalean ikusi den bezala, *Especial Case 1* marka ezarriko zaie, eta horrelakoen HAULak agertzean (*euskal pilota*) *Especial Case 3* eta *Especial Case1* erabiliko dira, adiera *usu* bat falta dela adierazi nahi delako.

[36] Kasu honetan, editoreak banatutako synset-zerrendan *partidu politiko* eta *partidu*, **biak** synset berean badaude (ikus 15. irudia), orduan, etiketatzaileak a., b. eta c. pausoak jarraituko ditu, baina agerpen horri synset horretako *partidu* variant-aren adiera-zenbakia emango dio, hau da, *partidu_2*.

Especial Case 3 bezala etiketatuko du, synset bat ematen saiatuko da[37] eta emandako synset-a apuntatuko du[38].

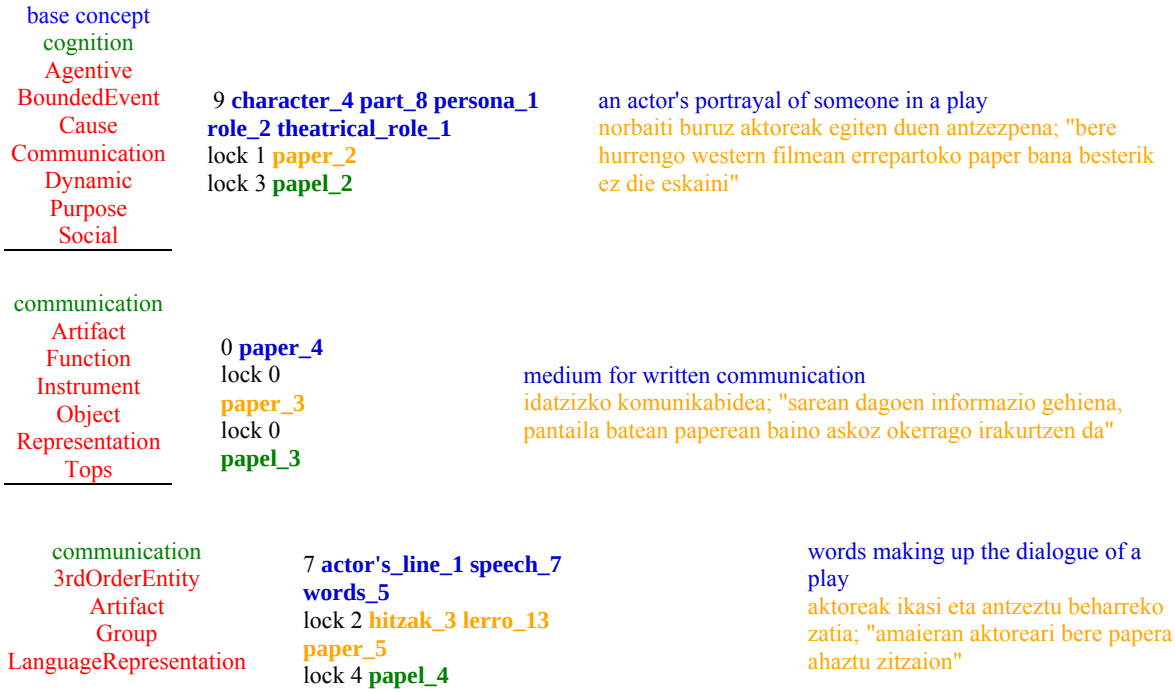


16. irudia

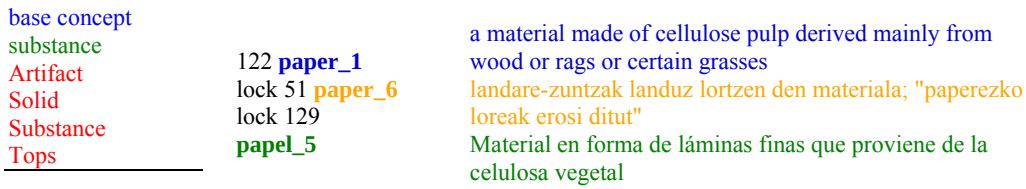
Oharra: gaur egun, oraindik ez dago guztiz adostuta HAULa zer den. Egun, IXA taldeko hainbat azpitalde HAULak aztertzen ari diren arren, oraindik definitzeko dago IXA taldeak –bere osotasunean– zer ikuspegiren arabera landuko dituen.

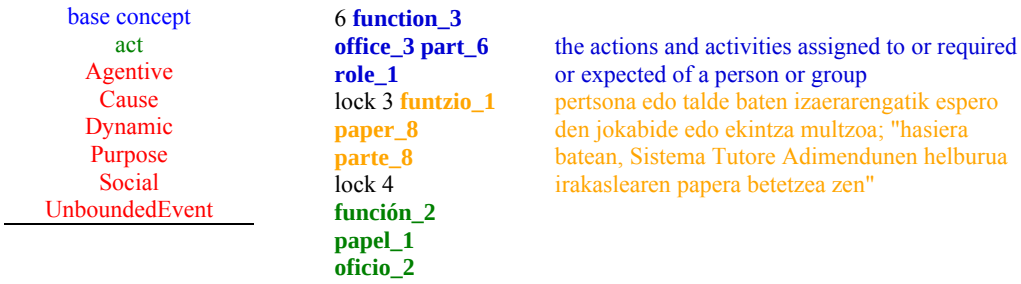
Bestalde, *Especiak Case 3* honek beste erabilera batzuk baditu: forma jokatuekin, eta diskurtsoaren adierazgailu gisa erabiltzen diren egitura horiekin.

❖ Gerta daiteke etiketatu beharreko hitza, **agerpenen batean, jokatua egoteagatik adiera ezberdina izatea**. Demagun, *paper* hitza etiketatu behar dela, EuskalWordNet-eko hurrengo bost synset-ekin:



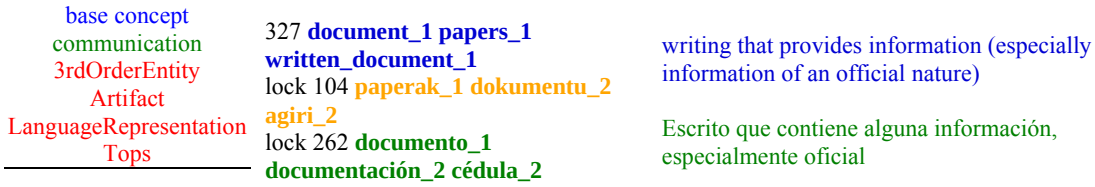
[37] Etiketatzaileak ez badu synset egokirik ikusten, eta *Especial Case 3* eta *Especial Case 1* markak batera ezartzeko beharra ikusten badu, hitz horren adiera *usu* bat falta denaren adierazgarri da.
[38] Ondoren, editoreak aztertu beharko du ea *partidu* eta *partidu politiko* gauza bera adierazteko erabil daitezkeen, hots, synset berean elkarrekin egon daitezkeen.





17. irudia

Hitz hau etiketatzean, etiketatzaileak corpusean "Beharrezkoak ziren paperak sinatu zituzten" etiketatu beharko balu, *Especial Case 3* marka erabiliko luke; *paper* forma pluraleko markarekin ez baitator bat goiko bost synset-ekin, hau da, *paperak* adiera berri bat ('dokumentu/agiri') adierazten du. Horrela, *paper* eta *paperak* hitzen adierak ezberdintzen dira; lehenengoak goiko bost synset-ak ditu, eta bigarrenak, aldiz, *paper* hitzak **singularrean** ez duen bestelako adiera bat du:



18. irudia

Hortaz, 'dokumentu' esanahia duenean, *paper* hitza **beti pluralean** agertzen da, eta horregatik, *Especial Case 3* bezala markatzen da, jokatua egoteagatik adiera ezberdina duelako.

Hona hemen, jokatua egoteagatik adiera ezberdina duen beste adibide bat:

- Indar handia du.
- Indarrez sartu eta biztanle guztiak hil egin zituzten

Lehenengo adibidean, *indar* hitzak 'ahalmen fisikoa' adierazten du, eta bigarrenean, aldiz, *indarrez* 'indarraren bidez'-edo. Hala, *indar* hitzak 'indarraren bidez' adierazteko

beti jokatua (-z) egon behar du, eta arrazoi horregatik, *Especial Case 3* bezala markatzen da[39].

Oharra: Flexioak adieran eragiten duen kasu hauetan, *Especial Case 3* markarekin batera, ez da beharrezkoa synset bat ematen saiatzea[40].

❖ **Diskurtsoaren adierazgailu gisa erabiltzen diren egitura horiek** (*hasiera batean, alde batetik,* eta abar) HAULen irizipide bera jarraituaz etiketatuko dira.

Oharra: Diskurtsoaren adierazgailu gisa erabiltzen diren egitura hauetan, *Especial Case 3* markarekin batera, ez da beharrezkoa synset bat ematen saiatzea.

5.2.1.4. Especial Case 4: Word is part of a Named Entity

Kasu berezi hau izen bereziak edo entitateak markatzeko erabiltzen da. Etiketatu beharreko hitza, agerpenen batean, izen berezi/entitate bat bada, edota izen berezi/entitate bat osatzen duen osagaietako bat bada, orduan, agerpen horri *Especial Case 4* marka emango zaio[41]. Adibidez, *kutxa* hitza etiketatzean, corpusean *Kutxa* (aurrezki-kutxaren izena) ager daiteke:

- Ezpain-margoak gordetzeko kutxak egin eta saltzen ditu.
- Argazkian, Josu Jon Imaz eta Fernando Spagnolo Kutxako lehendakaria.

Bigarren esaldian, etiketatu beharreko agerpena izen berezi/entitate bat denez, agerpen hori *Especial Case 4* bezala etiketatuko da.

Gerta liteke, hitz batek izen berezi/entitate bat adieraztea, baina idazlearen arabera, hitz hori corpusean letra larriz edota minuskulaz idatzirik egotea, esate baterako, *estatu*:

- Exekutiboak ustez erabateko kontrola du estatuko gainontzeko botereen gainean.
- Pirinio Atlantikoen " kultura askotarikoa " dela eta , Estatuak departamenduan Paue eta Baionako aglomerazioa ezberdintzen direla ikusirik...

Bi esaldi hauetan, *estatu* hitza batean letra larriz eta bestean xeheaz dagoenez, zalantza sor dezake etiketatzailearengan: dagokion synset-arekin etiketatu ala *Especial Case 4* bezala markatu? Honelako kasuak **aurreikusten direnean**, bileran komentatu/erabaki beharko dira etiketatzean koherentzia gordetzearren. Hala ere, oinarri

---

[39] Azken finean, lematizatzaileak horrelako formak ezberdintzen gai izango balitz, *paper/urte* eta *paperak/urteak* forma ezberdin gisa etiketatuko lirateke. Egun, lematizatzaileak horrelakoak harrapatzea zaila denez, *Especial Case* hau sortu behar izan da, eta hala, etorkizunean lematizatzaileak horrelako marketatik *ikasi* ahal izango du. Horregatik, *Iruzkinetarako taula* txostenean informazio hau txertatu beharko du.

[40] Horrelako kasuak, *paperak* eta *indarrez* bezalakoak, *Especial Case 3* eta *Especial Case 6* etiketatuko dira. Jokatuak izateagatik, adiera ezberdina izateaz gain, gainera, gaizki lematizatuak daudelako (ikus 5.2.1.6 atala).

[41] 3lb tresnan gorriz markatuta dauden hitzak – etiketatu beharreko hitzak, alegia – lematizatzailetik (EDBLtik) datoz. HAULekin gertatzen zen bezala, gaurko EDBL bertsioarekin, lematizatzaileak ezin ditu izen berezi/entitateak bereizi, horregatik sortu behar izan dugu *Especial Case* hau.

gisa *Egunkariako Estilo Liburua* erabiliko da. Bertan, behar den laguntza aurkituko ez balitz, hurrengo irizpidera joko da:

❖ Gure ustez, izen horrek entitate bati erreferentzia egiten badio, nahiz eta letra xeheaz egon corpusean, ***Especial Case 4*** bezala markatuko da[42].

Azkenik, zenbait kasutan izen berezi/entitate bat, aldi berean HAULa izan daiteke, hau da, agerpenen batean etiketatu beharreko hitza izen berezi/entitate bat osatzen duen osagaietako bat izan daiteke, **eta gainera**, HAUL baten osagaia ere izan daiteke. Demagun *herri* hitza etiketatu behar dela:

- Zenbait erakunde juridikoren bidez herriko biztanleen artean sortu da elkartasuna.
- ...garatze prozesuan dagoen Euskal Herriko entziklopedia elektronikoa.
- Herri Batasunak ere kondenatu egin zuen ekintza

Lehenengo esaldian ez bezala, bigarren eta hirugarren esaldietako *herri* hitza alde batetik, HAUL baten osagaiak dira (*Euskal+Herri; Herri+Batasuna*), eta bestetik, HAUL horiek izen berezi/entitate bat izendatzen dute. Horrela bada, horrelako kasuak ***Especial Case 3*** **eta** ***Especial Case 4*** bezala etiketatuko dira. Hala ere, hurrengo adibideen antzekoak topatuz gero, zalantzak sor daitezke:

- Errusiako Gobernua / Errusiako gobernua
- Errusiar Gobernua / Errusiar gobernua / Gobernu errusiarra / gobernu errusiarra
- Moskuko Gobernua / Moskuko gobernua
- Putin-en Gobernua / Putin-en gobernua

Horrelako zalantzak ebazteko, *Egunkariako Estilo Liburu*ra[43] joko da. Dena den, adibide batzuk oso zehatzak dira, eta ez dira *Egunkariako Estilo Liburu*an agertuko. Kasu hauetarako, irizpidea hurrengoa da[44]:

---

[42] Entitateek, oro har, ez dituzte ondorengo ezaugarriak hartzen: ez dira pluralean agertuko (*Hainbat **estatuetako** agintariek erabaki dute...*); ez dira nolakotasunezko adjektiboekin agertzen (***Estatu** anarkista bat osatzeko...*), ezta determinatzaile mugagabeekin ere (*Hainbat **estatuetako**...*).

[43] *Egunkariako Estilo Liburua* kontsultatzen dugula esaten dugunean, letra larri eta xeheei buruzko informazioaz ari gara; talde osoak gai honi serioski ekin arte, letra larriaz idatzi beharrekoak Izen Berezi/Entitateekin identifikatzeko erabakia hartu dugu. Adibidez, *sari* hitza etiketatzerakoan, *Egunkariako Estilo Liburuan Nobel saria, Iparragirre saria, Espainiako Literatura Sari Nazionala, Frantziako Sari Nagusia, Monakoko saria* aurkitu dugunez, letra larriz daudenak bere adiera, C3S eta C4S moduan markatzea erabaki dugu eta letra xehez daudenak bere adierarekin soilik. Zalantza *Frantziako saria / Frantziako Saria* nola etiketatu behar diren erabakitzerakoan etorri da eta *Frantziako* elementuak *sari* kontzeptua bere gain hartzen ez duenez, C3S eta C4S markak jartzea erabaki dugu. Kontuan hartu *sari* eta *gobernu* hitzen gaineko irizpideen desberdintasuna euren izaera semantiko ezberdinagatik izan dela.

[44] Kasu hauetarako irizpide batzuk erabaki dira, baina kontuan izan beharrekoa da hauek koherentzi bat mantentzeko hartuak izan direla, eta horregatik, behin-behineko moduan hartu behar direla talde osoak gai honi serioski ekin arte.

❖ HAULak entitate bati erreferentzia eginez gero, nahiz eta letra xeheaz egon, ***Especial Case 3*** **eta** ***Especial Case 4*** bezala etiketatuko dira (goiko lehenengo hiru adibideak[45]). Azken adibidearen kasuan, (Putin-en Gobernua / Putin-en gobernua), *gobernu* hitza ***Especial Case 4*** bezala markatuko da, baina *Putin-en gobernua* ez da HAUL gisa ulertuko[46]. Bestalde, *Espainiako nahiz Frantziako Gobernuak* estilo liburuan pluralak letra xehez jarri behar direla dio. Beraz, guk ez diegu ez HAUL eta ezta IZB jarri.

5.2.1.5. Especial Case 5: The tagger is strongly uncertain

Etiketatzaileak marka hau zalantzak dituenean erabiliko du, hau da, etiketatu beharreko hitza, agerpenen batean, zalantzazkoa bada (testuinguru faltagatik, adibidez), eta etiketatzaileak ez badaki zer synset-ekin markatu, orduan erabiliko du kasu berezi hau. Marka hau era ezberdinetan erabili daiteke:

a. **Etiketzaileak agerpen batekin zalantzak dituenean**, orduan, besterik gabe, *Especial Case 5* marka ezarriko dio.

b. **Etiketatzaileak agerpen bat synset-zerrendako synset bat izan daitekeela uste duenean, baina erabat ziur ez dagoenean**, agerpen horri *Especial Case 5* marka ezartzeaz gain, synset bat egokitu ahal izango dio, eta horrela, etiketatzaileak aukeratutako synset horrekin erabat ziur ez dagoela adieraziko du.

c. **Etiketatzaileak testuinguru faltagatik** (esaterako, egunkarietako izenburu[47] bat bakarrik azaltzen denean) etiketatu beharreko hitza zehazki zein synset duen argi ikusten ez badu, *Especial Case 5* markatuko du, nahiz eta fitxategi berean, aurreko edota ondorengo agerpenek argibideren bat ematen dutela pentsatu.

5.2.1.6. Especial Case 6: Word was improperly lemmatized or PoS-tagged

Etiketatzaileak etiketatu beharreko hitza, agerpenen batean, ez badator bat etiketatzen ari den hitzaren kategoriarekin edo gaizki lematizatua badago, orduan, *Especial case 6* bezala etiketatuko du. Adibidez, *etxe* hitza etiketatzean hurrengo esaldia bezalakoak aurki ditzakegu:

- Etxeko gazta fresko ikaragarria dastatzeko aukera izango duzue.

Kasu honetan, agerpen horri *Especial Case 6* marka emango zaio, *etxeko* forma ez delako *etxe (izena)+-ko (atzizkia)*, baizik eta, *etxeko* adjektiboa (‘etxean egindakoa’ adierazten duen adjektiboa). Lematizatzailea erratu egin da, nahiz eta itxuraz berdinak izan, kategoria ezberdina duten bi forma dira[48].

---

[45] *Moskuko Gobernua* metonimia gisa ulertu da, hau da, *Errusiako Gobernua* adierazteko beste modu bat dela erabaki da.

[46] Horrelako jabetza-esamoldeak (*Putin-en gobernua*) termino bera (*Errusiako Gobernua*) behin eta berriz ez errepikatzeko erabiltzen dira. Esamolde mota horiek HAULak ez kontsideratzea erabaki da.

[47] Etiketatu beharreko agerpena izenburu batetako hitz bat bada, edozein hitz bezala etiketatuko da.

[48] *Especial Case 6* diren kasu batzuk, aldi berean, *Especial Case 3* dira, hau da, jokatuak daudenez adiera ezberdin bat hartzen dutenez, *Especial Case 3* marka ere hartzen dute (adibidez, *etxeko, paperak, indarrez*…)

Beste adibide batzuk:

a. *Peru* ('herria') etiketatzean *Perurena* (izen berezia) agertzea corpusean:

   - Peru herrialde handia da.
   - Perurenak harria jaso zuen.

b. *Eraikin* hitza etiketatzean corpusean *edifizio* lema agertzea[49]:
   - Erakin horretan bizi naiz.
   - Edifizio hori harriz egina dago.

c. Corpusean beste hizkuntzetako hitzak agertzea: *a* ('hizkia') etiketatzean ingeseseko edo gaztelaniako *a* artikulua eta preposizioa lematizatzea corpusean:

   - a cup of tea
   - ir a comer

5.2.1.7. Especial Case 7: Word is wrongly used: misspelling or erderakada

Kontuan izan beharrekoa da, corpusean agertzen diren hitz, adiera eta erabilera guztiak ez dutela nahitaez zuzenak izan behar, akatsak egon daitezkeela. *Especial Case* honekin, hauek markatzen dira, hain zuzen ere. Hortaz, etiketatzaileak kasu berezi hau erabiliko du etiketatu beharreko hitza, agerpenen batean, gaizki erabilita edo erdarakada garbia bada. Beraz, marka hau ondorengo kasuetan erabili daiteke:

a. **Erderakak**: Gaztelaniaren kalkoa diren egitura ez zuzenak ager daitezke:

   - Joannes maiteari ilea hartu nahian ibili da.

   Ilea hitza etiketatzean, *ilea hartu* bezalako egituraren bat agertuko balitz, erderakada gisa – *tomar el pelo* – markatu beharko genuke, *Especial Case 7* gisa, alegia; euskaraz egitura hori ez baita zuzena, hori adierazteko *adarra jo* erabili beharko baikenuke.

b. **Errore ortografikoak**: Hau erabiliko dugu corpusean gaizki idatzitako hitzen bat agertuz gero, baina lematizatzaileak lema identifikatu badu: **iharduera jardueraren* ordez, **pake bakeren* ordez, **konpromezu konpromisoren* ordez, **Euskal Autonomi Elkarte Euskal Autonomi Erkidegoren* ordez, eta abar.

[49] Kasu hau berezia da, *edifizio* ez da forma estandarra, eta *eraikin* erabiltzea gomendatzen da. EDBLk informazio hau guztia jaso egiten du, eta lematizatzailea EDBLn oinarritzen denez, EDBLko informazio honek guztiak lematizatzailearen emaitzean eragina izan dezake; hau da, *edifizio* hitza forma ez-estandarra denez, lematizatzaileak EDBLn duen forma estandarraren lema egotzi egiten dio: *eraikin* lema, hain zuzen ere. Hala, lematizatzailearen errore gisa jotzen dugu, *edifizio* hitzaren lema, ez-estandarra izan arren, *edifizio* delako eta ez *eraikin*, hots, lema ezberdina dutelako. Honekin batera, eta 5.2.1.7 atalean aipatuko dugun bezala, *edifizio* hitza corpusean agertzean, ez zuzentzat hartuko dugu, honen ordez, euskaraz *eraikin* erabili behar delako.

Hona hemen beste adibide bat: kontuan izanda *ekialde* (gaztelaniako 'este' adierazteko) eta *Ekialde* ('Asia' adierazteko) – lehena minuskulaz eta bigarrena letra larriz– lematizatzaileak bereizi egin beharko lituzkeela, etiketatzailea minuskulaz idatzitako ***ekialde* etiketatzen** ari bada, *Especial Case 7* markatuko du baldin eta corpusean dagoen hitza **minuskulaz badago eta maiuskulaz duen adierari egiten badio erreferentzia**.

- Bart, ekialde urruneko biztanleak manifestazio handietan bildu ziren.

Aldiz, *ekialde* etiketatzean corpusean *Ekialde* agertuko balitz, eta alderantziz, *Ekialde* etiketatzean corpusean *ekialde* agertuko balitz, *Especial Case 6* jarri beharko da, izan ere, lematizatzaileak maiuskulaz eta minuskulaz idatzitakoak bereizi beharko lituzke, hauetako bakoitzak hiztegi sarrera aparteko bat baitute.

c. **Bestelako erroreak**: Hauek aurrekoetatik ezberdintzen diren akatsak dira. Esate baterako, 5.2.1.6 ataleko adibide bat *Especial Case 7* ere bada:

- Erakin horretan bizi naiz.
- Edifizio hori harriz egina dago.

Kasu hau berezia da, *edifizio* ez da forma estandarra, eta *eraikin* erabiltzea gomendatzen da. 5.2.1.6 atalean esan dugun bezala, kasu hau lematizatzailearen errore gisa (*Especial Case 6* gisa) jotzen dugu, *edifizio* hitzaren lema, ez-estandarra izan arren, *edifizio* delako eta ez *eraikin*. Honekin batera, *edifizio* hitza corpusean agertzean *Especial Case 7* marka ezarriko zaio, euskaraz ez-zuzena dela adierazteko, hain zuzen ere.

5.2.2. Bestelako arazo batzuk

Etiketatzean ohiko arazoak honako hauek dira:

1. Etiketatzen ari garen hitza gorriz azpimarratu beharrean ondokoren bat (orain arte hitzaren aurrekoa) azpimarratu izana izan ohi da 3lb tresnaren arazorik arruntena. Kasu hauetan, zein fitxategitan gertatu den jakinarazi behar zaio arduradun den informatikariari.

2. Beste arazo bat siuc01 edo siuc02 makinetan memoria asko eskatzen duten prozesuak daudenean gertatzen da; honelakoetan, oso geldiro joaten da eta etiketatzeko lana aspergarri bezain luzea gertatzen da; orduan, arazoaren berri emango zaio informatikariari, honek lehentasun mailak aldatuz etiketatzeari emango diolarik lehentasuna.

3. Bestalde, siuc02 makinan letra tipo bat falta da; beraz, fitxategi bat ireki behar den guztietan halabeharrez erantzun behar zaio letra tipoa aldatzeko galderari. Galdera erantzuten ez ibiltzeko eta lana sinpletzeko siuc01 makina erabiltzea gomendatzen da.

4. Fitxategiren batekin arazo informatikoren bat gertatuz gero, etiketatzaileak fitxategi horren izena apuntatu eta informatikariei bidaliko dio posta-elektronikoaren bitartez, (baita beste etiketatzaileari ere). Informatikaria izango da fitxategi hauek gordetzeko ardura duena.

5. Amaitzeko, 3lb interfazean sartzen garenean, ezker aldean agertzen diren hitzen artean baten batzuk laranjaz egon daitezke. Hasiera batean, bertan sartu eta etiketatzen ziren. Hala ere, egun, monosemikoen lanketaren ondorioz, izen asko eta asko laranjaz ager daitezke (ikus 7. atala), eta arrazoi horregatik, laranjaz dauden horiek ez etiketatzea erabaki da.

5.3. Etiketatu ondorengo eginbeharrak

1. Etiketatzaileak etiketatutako hitzen synset-ak itzulpena egingo du. Hitz bat etiketatu bezain laster egitea komeni da, izan ere, synset bakoitzaren nondik norakoak argiago izango ditu berriki etiketatu dituen horietan. Itzulpenak egiteko prozedura-txostena eskuliburu honen eranskinetan dago (ikus B atala).

2. Etiketatzaileek lana amaitu ondoren, etiketatutako hitzaren datuak (hitza bera, agerpen kopurua, synset kopurua, etiketatzeko eta itzultzeko behar izan duen denbora) fitxategi batean idatziko dituzte. Etiketatzaileek partekatzen duten fitxategia da hau.

3. Etiketatzaileek**,** etiketatze-lana amaitu ondoren, lan-taldeari jakinaraziko diote.

4. Etiketatzaileak adiera berriren bat (*Especial Case 1*) edo HAULen bat (*Especial Case 3*) etiketatu badu, honen berri eman behar du bileran, adiera berri hori zein den adieraziz.

5. Bileran, etiketatzean izan dituen zailtasunak eta hartu dituen erabakiak jakinaraziko dizkio epaileari, baliagarria izan baitaitezke honen lanerako.
6. Epaileak bere lana egin ondoren, glosaren bat zuzentzeko eskatuz gero, glosa etiketatzaileek aldatuko dute. Horretarako, epaileak aldaketak zeintzuk diren zehaztu beharko die etiketatzaileei.

6. Epaitze-lana

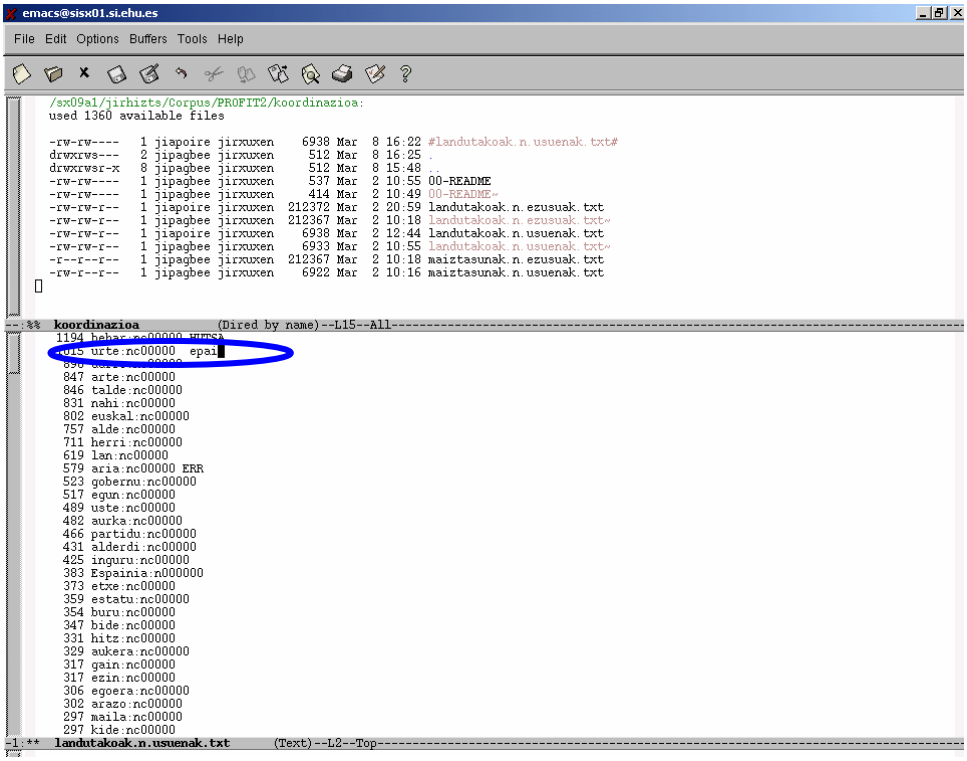
Epailearen lana, laburki esanda, etiketatzaileek etiketatutakoa konparatzea eta ezberdin dauden etiketen artean erabakitzea izango da. Eginbeharren azalpena ulerterrazagoa egiteko, eginbeharrak hurrengo fasetan banatuak izan dira:

a. Asteko bilera baino lehenagoko pausoak
b. Asteko bileran eginbeharrekoak
c. Epaitzean jarraitu beharreko pausoak
d. Epaitu ondoren jarraitu beharreko pausoak

6.1. Asteko bilera baino lehenagoko pausoak

Editoreak hitz bati **`etikOK`** marka jartzen dionean, etiketatzaileek hitz bakoitza etiketatzen amaitu dutela jakinaraziz (ikus 4.2 atala), eta etiketatzaileek glosak itzuli dituztela baieztatzen dutenean, epaileak hitza bilerarako prestatuko du. Prestaketa honek hainbat pauso ditu:

1. Aukeratutako hitzari **`epai`** marka jarriko dio dagokion fitxategian (ikus 1go puntua 4.1 atalean). 19. irudia 13.arekin konparatuz gero, *urte* hitzak zuen `etikOK` marka kendu da, epaileak `epai` marka jartzeko. Horrela, lantaldeko beste kideek badakite hitz honek marka hori duenean, epailea lantzen ari dela.



19. irudia

2. Epaileak, hitz baten agerpenak etiketatzen hasi baino lehen, editoreak banatutako hitz horren *fitxa* errepasatuko du. Bestalde, synset-zerrenda EuskalWordNet-etik inprimatua izan beharko du, azkeneko bertsioarekin lan egingo duela ziurtatzeko. Gainera, hurrengo atalean ikusiko dugun bezala, hau lagungarria izango da epailearentzat, etiketatzaileen etiketak alderatzen dituen fitxategiko synset-zenbakiak zein synset-i dagozkien jakiteko; adibidez, `<09534064>` synset-zenbakia *diru* hitzaren 2. synset-ari dagokio (*diru_2*).

3. Synset-zerrenda errepasatzean, glosaren batean akatsen bat aurkituko balu[50] (glosa ez dela ulertzen, akats ortografikoren bat, eta abar), bileran etiketatzaileei adierazi ahal izateko apuntatuko du.

6.1.1. Epaitzen hasi baino lehenagoko pausoak

Epaitu ahal izateko, lehenengo `etik1` eta `etik2`-k etiketatutakoa alderatu beharko du eta hori egin ahal izateko, programa hau exekutatu behar du (beti `PROFIT2` katalogoan egonda):

```
35-[siuc01 PROFIT2]% ./epailearenasortu `hitza´ `kategoria´
```

Demagun *diru* hitzaren alderaketa egin behar duela, orduan agindua hurrengoa litzateke:

```
36-[siuc01 PROFIT2]% ./epailearenasortu diru n
```

Return sakatu eta segundo batzuen ondoren, makinak alderaketaren emaitzarekin fitxategi berri bat sortzen du, eta fitxategi hori `Txostenak` karpetan gordetzen du. Karpeta honetara iristeko idatzi:

```
37-[siuc01 PROFIT2]% cd Txostenak
38-[siuc01 Txostenak]% ls
```

Eta bertan egongo da epaileak behar duen fitxategia, `diru.n.ita` bezala izendatua. Fitxategi hau ikusi ahal izateko *emacs* ireki behar du (*emacs* egikaritzeko ikus 4.1 atala). Fitxategi horren itxura ondorengoa izango da:

a. Batetik, etiketatzaile bakoitzak etiketatutako kopuruak adierazten dira. Adibidez, 1. etiketatzaileak `<09639711>` synset-a 91 agerpenetan erabili da; 51 agerpenetan `09639711` synset-a eta *Especial Case 3* (`C3S`)[51] marka batera, `<09642587>` synset-a 2 agerpenetan, eta abar.

```
1. etiketatzailea
-----------------
   Erantzunak:
   "              " 2
   "       09534064" 2
   "   09534064,C3S" 1
   "       09639711" 91
   "   09639711,C3S" 51
   "       09640689" 10
```

[50] Itzulpenak egiteko prozedura-txostena eskuliburu honen eranskinetan dago (ikus B atala).
[51] *Especial Case*-ak makinetan laburdura batzuekin agertzen dira: "C - *Especial Case*-aren zenbakia - S". Hortaz, *Especial Case 3*-a *C3S* laburtuko da, eta *Especial Case 5*-a *C5S*.

```
  "   09640689,C3S" 5
  "       09642587" 2
  "   09642587,C3S" 1
  "            C1S" 1
           Guztira: 166

 Erantzunak adieraka:
  "       09534064" 2
  "       09639711" 116
  "       09640689" 12
  "       09642587" 2
  "            C1S" 1
  "            C3S" 29
           Guztira: 164

 Adiera kopurua erantzuneko:
       <0:2>    <1:106>    <2:58>
        Batazbeste:  1.34 ( 1.35 hutsak kenduta)

2. etiketatzailea
-----------------
   Erantzunak:
   "               " 1
   "       09534064" 1
   "09534064,09639711" 4
   "       09639711" 99
   "   09639711,C3S" 45
   "       09640689" 11
   "   09640689,C3S" 4
   "       09642587" 1
            Guztira: 166

   Erantzunak adieraka:
   "       09534064" 3
   "       09639711" 123
   "       09640689" 13
   "       09642587" 1
   "            C3S" 24
            Guztira: 165
```

b. Bestetik, bi etiketatzaileen arteko adostasuna adierazten duten kopuruak daude. Kopuru hauetako batzuk adiera kontuan hartu gabe lortutakoak dira (ITA)[52], eta beste batzuk, aldiz, adiera kontuan hartuta (KAPPA):

   - Bi etiketatzaileek etiketatutako agerpenen kopurua (`ITA replies by both`)
   - Agerpenen bat hutsik geldituz gero, hori ere zehaztu egiten da (`ITA empty by one`)
   - Zenbateko adostasuna egon den etiketatzaileen artean (`ITA total agreement`)

[52] Hau Inter-tagger agreement (ITA) bezala ezagutzen da.

- *Especial case 6*, *Especial case 5*, *Especial case 7*-a eta etiketa gabe geratu diren agerpenak kontuan izan gabe, etiketatutako agerpenen kopurua zehazten da (`ITA good replies (non empty, non C6S, non C5S)`)

```
Adostasuna
-----------
ITA replies by both:    166
ITA    empty by one:    3
ITA    total agreement: 139 out of 163 (85.3%)
ITA good replies (non empty, non C5S/C6S/C7S):  163
ITA agreement on at least one sense:            157 out of 163 (96.3%)
Kappa: 0.92     (Adierak: 6 Pa: 0.9632 Pe: 0.5620)
```

c. Ondoren, bi etiketatzaileen arteko desadostasuna matrize batean adierazita dator (`Nahasmendu matrizea`). Bertan, synset-en arteko nahasketa eta beraien kopurua zehaztuta dago; esate baterako, 1. etiketatzaileak[53] `<09534064>` synset-a bezala etiketatu dituen 2 agerpen, 2. etiketatzaileak `<09639711>` bezala etiketatu ditu (`<09534064> <09639711>: 2`).

```
Nahasmendu matrizea
--------------------
  <09534064,C3S> <09639711,C3S>: 1
  <09534064> <09639711>: 2
  <09639711> <09534064>: 1
  <09639711,C3S> <09639711>: 8
  <09639711> <09639711,C3S>: 1
  <09639711,C3S> <09640689,C3S>: 1
  <09639711> <09534064,09639711>: 4
  <09640689,C3S> <09640689>: 2
  <09640689> <09639711>: 1
  <09642587,C3S> <09639711,C3S>: 1
  <09642587> <09639711>: 1
  <C1S> <09639711>: 1
```

d. Azkenik, nahasmendu matrizean adierazitako desadostasun horiek, zein fitxategitan aurkituko diren zehazten da. Adibidez, `eebs.450322049.sat.xml-1` fitxategian, 1go etiketatzaileak *diru* `<09639711,C3S>` bezala etiketatu du, eta 2.ak aldiz, `<09639711>` bezala. Eta horrela datoz zerrendatuak ados ez dauden fitxategi guztiak:

```
Desadostasunak
----------------
eebs.450322049.sat.xml-1
1:<09639711,C3S> 2:<09639711>
eebs.450740229.sat.xml-10
1:<09639711> 2:<09534064>
eebs.450740229.sat.xml-5
1:<09642587,C3S> 2:<09639711,C3S>
eebs.450740229.sat.xml-7
1:<09639711> 2:<09534064,09639711>
eebs.450740229.sat.xml-8
1:<09639711> 2:<09534064,09639711>
eebs.451222489.sat.xml-4
1:<09639711> 2:<09534064,09639711>
```

[53] Ezkerreko datuak beti 1. etiketatzaileari dagozkio, eta eskuinekoak 2. etiketatzaileari

Fitxategi honi esker, epaileak etiketatze-lanetan egon diren arazoak ikusiko ditu. Horretarako, gomendagarria da etiketatzaileen lanaren alderaketa aurrean izatea (inprimatzeko ikonoa ematearekin nahikoa da).

Bi etiketatzaileen lanaren alderaketa egiterakoan kontutan hartu behar dituenak:

1. Etiketatzaileek adiera berririk markatu duten begiratu (*Especial case 1*).
2. Adostasun maila baxua bada, desadostasuna sortzen duten adierak zeintzuk diren ikusi.
3. Glosak ondo ulertzen diren begiratu.

Asteko bilerarako ahal den informazio gehiena jaso behar du epaileak; beraz, bilera honen aurretik epaituko ez badu ere, gomendagarria da informazio hori lortzeko eskura dituen baliabide guztiak erabiltzea, adibidez aurrerago esplikatuko dugun `kwic` agindua.

6.2. Asteko bileran eginbeharrekoak

1. Etiketatutako azken hitzen emaitzak aurrean dituela, etiketatzaileen zalantzak eta iruzkinak jaso eta epaitzeko lagungarri izango zaion informazioa bilduko du (*Iruzkinetarako taula* txostenean). Gero, txosten horren azkeneko bertsioa web orrian[54] zintzilikatu beharko du .

2. Etiketatzailea, epailea eta editorea asteroko bileran elkartuko dira eta editoreak etiketatzaileari etiketatu beharreko hitzak banatuko dizkio (ikus 2.3 atala). Bilera horretan epaileak editoreak planteatzen dituen synset-ak ulertzen/adosten saiatuko da.

3. Bilera horretan bertan, aurreko astean epaitutako hitzei buruzko emaitzak/zalantzak/erabakiak jakinaraziko ditu.

6.3. Epaitzean jarraitu beharreko pausoak

1. Bilerarako prestaketa atalean aipatutakoari esker, epaileak etiketatzaileek synset-ekin izan duten nahasketa ikus/uler dezake, eta ondorioz, erabaki zein synset dagokion agerpen bakoitzari. ITA fitxategiko datuei esker epaileak epaitze-lana ikuspegi ezberdinen arabera landu dezake :

   - Batetik, epaileak **gehien errepikatzen diren fenomenoak batera landu ditzake (eta hauxe da gomendagarriena).** Adibidez, *gezi*ren kasuan, `<02533651>` eta *Especial Case 4*-ren arteko nahasketa askotan gertatu dela ikus dezake ITA fitxategian, eta ondorioz, bi etiketa horiek daramatzaten agerpen guztiak aztertzea komeni da, etiketen arteko nahasketa noiz, eta zergatik gertatu den ulertzeko eta gero, erabaki bat hartu ahal izateko.

[54] http://ixa.si.ehu.es/Ixa/Azpitaldeak/Lexikoa%20eta%20semantika

- Bigarrenik, **behin bakarrik gertatu diren nahasketak azter ditzake**. Nahasketa horiek konpontzeko zuzenean agerpen horiek dauden fitxategietara jo beharko da, agerpen zehatz hori zein adiera duen erabakitzeko (adibidez, agerpen bakarra dago, non *Especial Case 1* eta *Especial Case 3* nahastu diren). Fitxategietara jotzeko eta hauetan erabakitako aldaketak egiteko, epaileak 3lb tresna erabili beharko du. Horretarako, ikus 3. atala.

- Azkenik, **hitz batean adostasun maila baxua dagoela ikus dezake.** Horrelakoetan, epaileak berak zalantzak baditu ere, zalantza horiek apuntatuko ditu eta asteko bilerara eramango ditu, denen artean konponbideren bat aurkitzen saiatzeko. Zailtasunak egongo balira, editoreak hitz hori berriro landu beharko luke.

2. Epaileak etiketatzaileek *Especial Case 1* bezala markatutako agerpenak errepasatuko ditu, eta hala direla baderitzo, hurrengo bileran editoreari jakinarazteko apuntatuko ditu.

3. *Especial Case 3* edo HAUL bezala markatutako agerpenekin desadostasunik balego, epaileak kasu hauek alde batera utziko ditu[55], baldin eta argi eta garbi HAUL direla baderitzo.

4. Bere ustez HAULak direnak apuntatuko ditu, eta hurrengo bileran editoreari emango dizkio EuskalWordNet-en sartzeko.

5. **Kontuz**! *Especial Case 3* etiketa flexioak adieran eragiten duen kasuetan ere erabiltzen da; kasu hauetan desadostasunik balego, epaileak agerpen horiek epaitu beharko ditu.

6. Askotan epaitzeko garaian (HAULak eta adiera berriak diren ala ez erabakitzeko adibidez), epaileak hiztegiak eta bestelako baliabideak erabili beharko ditu (ikus 4.1 atala).

7. Epaitzen ari den bitartean sortu zaizkion zalantza pertinente guztiak apuntatuko ditu lan-taldearekin egingo den hurrengo bileran aztertzeko.

Epaitzen hasteko irizpideak hartu aurretik epaileak adibide mordoska bat begiratu nahiko balu, 3lb tresna erabili beharrean `kwic` aginduaren bidez egin dezake. Agindu hau `PROFIT2`-n exekutatzen da eta hitz baten agerpen guztiak bere datu guztiekin eman ditzake (ikus 3.3 atala).

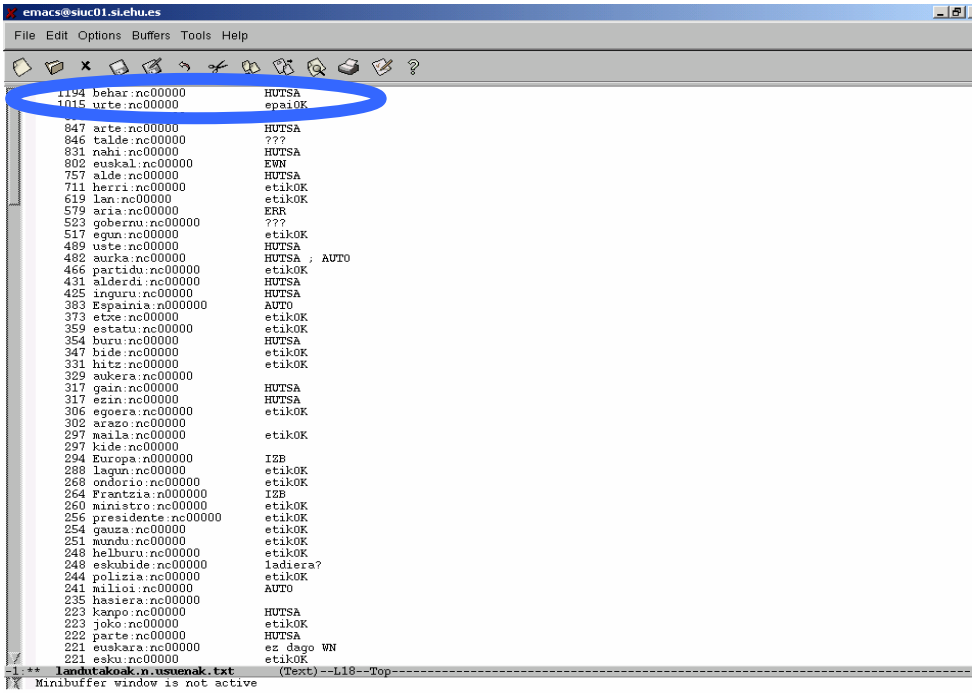
[55] HAULen orrazketarako irizpideak hurrengo urrats batean egingo direnez, eta IXA taldeko beste azpitaldeekin koordinatu beharreko zerbait denez, oraingoz HAUL bezala markatutakoa gorde egingo dugu.

6.4. Epaitze-lana amaitzean eginbeharrekoak

1. Hitz bakoitzarekin igarotzen duen denbora apuntatuko du.

2. Epaitzerakoan hartutako erabakiak eta hauetarako erabili dituen irizpideak *Iruzkinetarako taula*n zehaztu beharko ditu (grisez). Besteak beste, taula honetan, beti plurala hartzen duten izenen berri eman beharko du[56] (*Jokoak = Joko Olinpiarrak*).

3. Hitz baten epaitze-lana amaitzean, epaileak **`epaiOK`** marka jarriko dio dagokion fitxategian (ikus 4.1 atalean). 20. irudia 19.arekin konparatuz gero, *urte* hitzak zuen `epai` marka kendu da, epaileak `epaiOK` marka jartzeko. Horrela, badakigu hitz honek marka hori duenean, epaileak bere lana amaitu duela.



20. irudia

4. Hitz baten epaitze-lana amaitzean, epaileak `epaiOKC1S` marka jarriko dio dagokion fitxategian (ikus 4.1 atalean), **baldin eta epaitutako hitzak adiera berriren bat duen**, hau da, epaileak bere lana amaitu duen arren, hitz horrek adiera berria(k) d(it)u EuskalWordNet-en txertateko. Hitz horiek adiera berri hori EuskalWordNet-en txertatu arte, `epaiOKC1S`[57] eramango dute. Txertatu ondoren, epaileak hitz horren *Especial Case 1* guztien ordez, sortutako synset berri horrekin markatuko ditu. Eta azkenik, `epaiOKC1S` marka `epaiOK` markagatik aldatuko du (epaitze-lana guztiz bukatua dagoen markagatik, alegia).

[56] Ikus 5.2.1.3 atala.
[57] C eranskinean hitzei ematen zaizkien marka guztiak datoz zerrendatuak.

5. Hurrengo bileran, editoreari corpusean agertu diren adiera berriak eta ziurtzat dituen HAULak jakinaraziko dizkio

6. Epaileak synset baten esanahia eragingo lukeen aldaketaren bat egingo balu, glosaren itzulpena eta adibideak begiratu beharko lituzke, eta berak erabakitako esanahiarekin bat etorriko ez balira, etiketatzaileei glosak zuzentzea eskatuko die, akatsa non dagoen adieraziz.

7. Epaileak synset bat kentzea erabakitzen duenean, EuskalWordNet ikutu aurretik, kontutan izan behar du 3lb tresna ere aldatu egingo dela. Beraz, synset horrekin markatuta dauden agerpenak hutsik geratuko dira eta beste adiera edo/eta *Especial Case* bat zutenak adiera hori edo/eta *Especial Case* hori bakarrik izango dute. Hala ere, aldaketa egin eta lehenengo aldian `./ireki` agindua erabiliz gero, EuskalWordNet-etik ezabatu den synset-arekin markatuta zeuden adibideak ireki ahal izango dira. Behin fitxategia ikusi eta itxita, synset horren marka galdu egingo da. Hala ere, synset bat ezabatu baino lehen, editoreak eta epaileak koordinatuta egon beharko dute hau guztia eragoztearren.

7. Monosemikoak lantzen

5. atalean aipatu bezala, eskuzko etiketatzea hitz polisemikoekin bakarrik egiten da, hau da, EuskalWordNet-en hitz batek synset bat baino gehiago duenean. Monosemikoak diren hitz guztiak, hurrengo fase baterako uzten dira, zeinetan hitza monosemiko horiek guztiak automatikoki etiketatuko diren.

Hortaz, hitzak monosemikoak ala polisemikoak diren, EuskalWordNet-en dituzten synset-en arabera erabakitzen da. Hala ere, atal honetan azalduko den bezala, honek ez du ziurtatzen hitz horrek adiera bakarra duenik; hau da, EuskalWordNet ingeleseko WordNet-en oinarrituta garatzen ari da, eta nahiz eta lan honetan lauzpabost urte eraman, egun oraindik landugabeko edo orraztu beharreko synset-ak daude. EuSemCor proiektu honen helburuetako bat horixe da, hain zuzen ere: euskal hitz gehienen adiera arruntenak EuskalWordNet-en daudela ziurtatzea (ikus 1go atala).

Hala, hitz monosemikoak automatikoki tratatu baino lehen, benetan monosemikoak direla egiaztatu behar da, hots, EuskalWordNet-en beste synset-ik ez dutela. Atal honetan lan horren berri pausoz-pauso emango dugu, erabakitako hainbat irizpiderekin batera.

7.1. Hitz monosemikoa aukeratu

Aurreko ataletan (ikus 2.2 atala adibidez), behin baino gehiagotan aipatu dugu corpuseko hitzak zerrendatuak daudela corpusean duten maiztasunaren arabera (maiztasun handienetik txikienera). Berez, zerrenda hau bitan banatuta dago: zerrenda batek corpusean usuenak diren hitzak izango ditu (`landutakoak.n.usuenak.txt`), eta besteak, berriz, ez usuak direnak (`landutakoak.n.ezusuak.txt`). Hauek `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan (hauetara iristeko ikus 4.1 ataleko azalpena) daude. Hasiera batean, bi zerrendetatik hartutako izenak lantzen/etiketatzen baziren ere, gaur egun, usuenen zerrenda bakarrik baliatzen da, hau bukatu arte behintzat. Dena den, monosemikoak lantzerakoan ez usuak diren hitz horiekin hasi da, monosemiko kopuru handiena zerrenda honetan egongo dela susmatzen baita. Horrela bada, `landutakoak.n.ezusuak.txt` fitxategiko hitz monosemikoak orraztuko dira lehendabizi, eta ondoren, `landutakoak.n.usuenak.txt` fitxategikoak.

Fitxategi hauetako monosemikoak `maiztasunak.n.ezusuak.adiera1.txt` eta `maiztasunak.n.usuenak.adiera1.txt` fitxategietan daude `ADIERABAK` bezala markatuak. Fitxategi horiek `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan daude (hauetara iristeko azalpena 4.1 atalean dago).

7.2. EuskalWordNet kontsultatu

`ADIERABAK` markak hitz horrek EuskalWordNet-en synset bakarra duela adierazten du, eta ataza honen bidez, hori egiaztatu nahi da, hain zuzen ere. Hori egiteko, lehenengo, zerrendako hitz monosemiko hori EuskalWordNet-en kontsultatuko da, bertan adierazten den kontzeptua zein den jakiteko/ulertzeko.

7.3. *Hitz monosemikoak diren egiaztatu hiztegiak baliatuz*

Behin kontzeptua ezagututa eta ulertuta, monosemikoa den ala ez egiaztatu behar da. Horretarako, hiztegiak baliatuko dira:

- ***Elhuyar Hiztegi Txikia*** (paperean)
- *Elhuyar Hiztegia* (http://www1.euskadi.net/hizt_el/)
- *Euskal Hiztegi Modernoa*
- *Euskal Hiztegia*
- *Hiztegixa* (http://ixa2.si.ehu.es/hiztegixa/)

Elhuyar Hiztegi Txikia hartuko da oinarri gisa. Oinarrizko hiztegi bat izaki, bertan euskarako hitzen adiera arruntenak datozelako, eta helburua EuskalWordNet-ek, gutxienez, bertan dauden adiera horiek guztiak izatea delako. Hala ere, gerta daiteke hiztegi honetan, adiera bat azaltzeko definizio edo adibiderik ez egotea, eta horrelakoetan, beste euskarako hiztegi elebakar/elebidunetara jo behar da. Hiztegietan begiratuta, hurrengo kasuak gerta daitezke:

a. EuskalWordNet-en synset bakarra zuenak, hiztegian ere adiera bakarra izatea, eta gainera, adiera hori EuskalWordNet-eko synset-arekin bateragarria izatea**.**

b. EuskalWordNet-en synset bakarra zuenak, hiztegian ere adiera bakarra izatea, baina, adiera hori EuskalWordNet-eko synset-arekin guztiz bateragarria ez izatea[58]. Demagun, *probintzia* hitza aztertu nahi dela. EuskalWordNet-en hurrengo synset-a du:

administration- -geography- base concept location Natural Object Part Place Tops	113 **province_1** **state_2** lock 35 **estatu_5** **probintzia_1** lock 153 **estado_5**	the territory occupied by one of the constituent administrative districts of a nation Estatu Batuak bezalako zenbait federaziotan, politikoki eta administratiboki zatitzen den lurralde bakoitza; "estatu#bere estatua hegoaldean dago" Cada uno de los territorios en que se dividen, política y administrativamente, algunas federaciones, como

21. irudia

Eta *Euskal Hiztegi Modernoa*k, berriz, definizio hau ematen du:

Hitza: probintzia
~~~~~~~~~~~~~~~~~~~~~~~~~
iz. A1) "Hainbat estatutako banaketa administratiboa, eskuarki botere zentralaren mende dagoena"

[58] Bi adierak bateragarria diren ala ez jakiteko, EuskalWordNet eta hiztegi elebakarretako definizioak parekatu behar dira.
~~~~~~~~~~~~~~~~~~~~~~~~~

Bi adiera hauek ez daude maila berean, EuskalWordNet-ekoa zehatzagoa baita. Adibidea hauen kasuan, editoreari ohartarazi behar zaio, geroago ikusiko dugun bezala[59].

c. **EuskalWordNet-en synset bakarra zuenak, hiztegian beste adiera gehiago izatea.** Adibidez, *droga* hitzak *Elhuyar Hiztegi Txikian* bi adiera ditu:

-pharmacy-
base concept
artifact
Artifact
Substance
Tops

657 **drug_1** something that is used as a medicine or narcotic
lock 213 **droga_1**
lock 275 **droga_1** Sustancia utilizada como medicamento o narcótico

Elhuyar hiztegia

euskara gaztelania
castellano vasco

droga

Adierak

1 iz.

Droga

Droga bigunak eta gogorrak: drogas blandas y duras
Garai bateko sendagileek erabiltzen zituzten drogak: las drogas que hace tiempo empleaban los medicos

2 iz. B

Disputa, riña, reyerta

Goiko andre-gizonak beti drogan dabiltza: el matrimonio de arriba siempre está de riña

22. irudia

Hiztegiei esker, hitz hori monosemikoa izan daitekeen ala ez jakin dezakegu, baina azkeneko erabakia hartzeko, corpuseko (EuSemCor-eko) agerpenak beharrezkoak izango dira (ikus 7.4 atala).

[59] Horrelakoei **AUTOC1S** marka ezarriko zaie (ikus 7.6.2 atala).

7.4. Hitz monosemikoak diren egiaztatu corpusak baliatuz

Hitz monosemikoa den ala ez hiztegietan egiaztatu ondoren, hitz horren agerpenak aztertuko dira corpusean, eta hurrengo kasuak gerta daitezke:

a. EuskalWordNet-en eta hiztegietan monosemikoa dena, corpuseko agerpen guztietan ere halaxe izatea, eta gainera, adiera hori EuskalWordNet eta hiztegiko adierarekin bateragarria izatea**:** Kasu honetan, lantzen ari garen hitzari `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan[60] dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTO`**[61] marka jarriko zaio. Honekin, hitz hori monosemikoa dela eta automatikoki landu daitekeela adierazten da.

b. EuskalWordNet-en eta hiztegietan monosemikoa izan arren (*burtsa* 'erakunde' gisa, adibidez), corpusean adiera hori ez den beste adiera berriren batekin agertzea (corpusean, erakundeaz gain, 'eraikuntza' ere ageri da: Bilboko Olabarri kalean kasinoa eta burtsa elkarren alboan). Horrelakoetan, lantzen ari garen hitzari (esaterako, *burtsa*) `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTOC1S`** marka jarriko zaio. Honekin, hitz hori **polisemikoa** dela eta **automatikoki landu ezin daitekeela** adierazten da. Hortaz, editoreak hitz hori landu beharko du EuskalWordNet-en lehenengo, gero etiketatzaileek hitz hori etiketatu ahal izateko. Bestalde, corpusean agertutako adiera berri horren/horien berri emateko **txostentxo bat idatziko da**[62].

c. EuskalWordNet-en synset bakarra zuenak (7.3 atalean aipatutako, *droga*, adibidez), hiztegian adiera gehiago izatea, baina corpuseko agerpenetan adiera horietako BAKARRA agertzea: corpusean *droga*ren lehenengo adiera bakarrik agertu da. Horrelakoetan, lantzen ari garen hitzari (esaterako, *droga*) `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTOC1S`** marka jarriko zaio. Honekin, hitz hori **polisemikoa** dela adierazten da. Hala ere, **corpuseko agerpen guztiak adiera bakarrari dagozkionez, automatikoki etiketatu daiteke**. Hortaz, nahiz eta editoreak hitz hori EuskalWordNet-en landu beharko duen, gero etiketatzaileek **ez** dute hitz hori etiketatu beharko. Horretarako, beharrezkoa da fenomeno hau beraien **txostenean azaltzea.**

[60] Katalogo honetara iristeko azalpena 4.1 atalean dago.
[61] C eranskinean hitzei ematen zaizkien marka guztiak datoz zerrendatuak.
[62] Corpuseko adiera kopurua hiztegietakoa baino handiagoa izatea, ez da batere arrunta. Kasu hauetan, segur aski, hitza gaizki erabili da (ikus 7.6.5 atala) edota polisemia erregularraren (ikus 4.2.1 eta 7.6.1 atalak) eraginez sortutako adieraren bat da.

7.5. Glosak

5.3 atalean azaldu dugun bezala, etiketatzaileak etiketatutako hitzen synset-en itzulpena egingo du. Itzulpenak egiteko prozedura-txostena eskuliburu honen eranskinetan dago (ikus B atala).

Monosemikoak diren synset horiek ere itzuliko dira, monosemikoak direla egiaztatu ondoren. Horrela bada, `landutakoak.n.ezusuak.txt` eta `landutakoak.n.usuenak.txt` fitxategian **`AUTO`** marka daramaten hitzen glosak euskaratuko dira EuskalWordNet-en.

7.6. Zenbait arazo

Hasiera batean, ataza hau erraza bazirudien ere, behin lanean hasita, zenbait arazori aurre egiteko, irizpide batzuk sortu ziren.

7.6.1. Polisemia erregularra

4.2.1 atalean aipatu dugun bezala, corpus bat etiketatzean *polisemia erregularra* fenomeno lexiko-semantikoa maiz ikus daiteke. Nahiz eta oso fenomeno arrunta izan, hiztegiek ezin dute fenomeno hau jaso, adiera horiek testuinguruaren baitan baitaude. EuskalWordNet-en ere horrelako hutsuneak arruntak dira, eta EuSemCor etiketatzean, maiz agertu izan dira horrelako adibideak. Hala ere, proiektu honen lehenengo fase honetan, polisemia erregularra alde batera utziko da, hurrengo urrats batean lantzeko asmoarekin.

Hortaz, hitz monosemikoak aztertzean ere horrelako kasuak alde batera utziko dira. Demagun, *Brasil* hitza aztertu behar dela, eta hau EuskalWordNet-en eta hiztegien arabera monosemikoa dela:

23. irudia

Nahiz eta hiztegietan monosemikoa izan, corpuseko agerpenetan bestelako adiera batzuk ager daitezke (Brasilek irabazi zuen txapelketa, adibidez). Kasu hau polisemia erregularreko adibide garbia da, eta testuinguruaren baitan daudenez, horrelakoak hiztegietan adieraztea zaila da. Arrazoi horrengatik, **monosemikotzat** joko ditugu. Hortaz, horrelakoetan, lantzen ari garen hitzari (esaterako, *Brasil*) `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan[63] dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTO`** marka jarriko zaio.

7.6.2. Adierak bateragarriak ez izatea

7.3 atalean esan bezala, batzuetan EuskalWordNet-en synset-a ez da beti hiztegiko adierarekin guztiz bateragarria (adibidez, euskarako *probintzia* eta ingeleseko *province*). Azaldu bezala, bi adiera maila berean ez daudenean, editoreari ohartarazi egingo zaio hauei `AUTOC1S` ezarriaz.

7.6.3. HAULak

Hitz bat nahiz eta monosemikoa izan –bai EuskalWordNet-en, bai hiztegietan eta bai corpusean–, corpusean HAULa gisa ager daiteke, hau da, corpus osoan adiera bakarrarekin agertu da, baina zenbait agerpenetan hitz hori HAUL baten osagai bat da[64]. Adibidez, *zin* hitz monosemikoa izan arren, corpusean *zin egin* bezalakoak ugaritan agertzen dira. Horrelakoetan:

a. HAUL diren agerpen horiek *Especial Case 3* bezala **etiketatuko** dira 3lb erabiliaz (5.2.1.3 ataleko irizpideak jarraituta).

b. HAULak etiketatu ondoren, `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan[65] dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTO`** marka jarriko zaio.

7.6.4. Izen bereziak eta entitateak

HAULen kasuaren antzekoa da honako hau, hots, hitz bat nahiz eta monosemikoa izan –bai EuskalWordNet-en, bai hiztegietan eta bai corpusean–, corpusean izen berezi/entitate gisa ager daiteke, hau da, corpus osoan adiera bakarrarekin agertu da, baina zenbait agerpenetan hitz hori izen berezi/entitate bat da[66]. Adibidez, *aberri* hitz monosemikoa izan arren, corpusean *Aberri* izen berezi/entitatea agertu da (Aberria ala hil). Horrelakoetan:

[63] Katalogo honetara iristeko azalpena 4.1 atalean dago.
[64] HAULei buruz 5.2.1.3 atalean mintzatu gara.
[65] Katalogo honetara iristeko azalpena 4.1 atalean dago.
[66] Entitate eta izen bereziei buruz 5.2.1.4 atalean mintzatu gara.

a. Izen bereziak/Entitateak diren agerpen horiek *Especial Case 4* bezala **etiketatuko** dira 3lb erabiliaz (5.2.1.4 ataleko irizpideak jarraituta).

b. HAULak etiketatu ondoren, `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan[67] dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTO`** marka jarriko zaio.

7.6.5. Lematizazio-erroreak

Aurreko bi ataletan aipatutakoarekin harremanetan dago atal hau ere, hots, hitz bat nahiz eta monosemikoa izan –bai EuskalWordNet-en, bai hiztegietan eta bai corpusean–, corpusean gaizki lematizatua ager daiteke, hau da, corpus osoan adiera bakarrarekin agertu da, baina zenbait agerpenetan hitz hori gaizki lematizatua dago, beraz, lematizatzailearen errore bat da[68]. Esate baterako, *estatua* ('eskultura') hitzaren agerpenen artean, *estatu+a* ('nazio') bezalakoak agertzen dira. Horrelakoetan:

a. Lematizazio erroreak diren agerpen horiek *Especial Case 6* bezala **etiketatuko** dira 3lb erabiliaz (5.2.1.6 ataleko irizpideak jarraituta).

b. Lematizazio erroreak etiketatu ondoren, `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan[69] dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`AUTO`** marka jarriko zaio.

Bestalde, gerta daiteke zerrendako hitza bat ez etortzea corpuseko agerpen bat berarekin ere, hau da, agerpen guztiak lematizazio erroreak izatea. Esate baterako, *ira* ('landarea' adierarekin), corpusean dituen agerpen guztiek *IRA* talde armatuari erreferentzia egiten diote. Horrelakoetan, `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan dagokion fitxategian (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`ERR`** marka jarriko zaio.

7.6.6. Bestelako erroreak eta erdarakadak

Hitz bat nahiz eta monosemikoa izan –bai EuskalWordNet-en, bai hiztegietan eta bai corpusean–, corpusean erabilera okerra eman zaio[70]. Adibidez, *subjektu* hitz monosemikoa izan arren, corpusean hurrengo erabilerarekin agertzen da: Gure lan honetan, subjektuak afasiko bilakatu aurretik elebidunak ziren. Adibide, honetan *subjektu* hitza erdarakada bat da, honekin gaztelaniako *sujeto* hitza adierazi nahi baita. Euskaraz, testuinguru honetarako *pertsona* edo *lagun* hitzak erabili beharko lirateke: Gure lan honetan, pertsona horiek afasiko bilakatu aurretik elebidunak ziren. Horrelakoetan:

[67] Katalogo honetara iristeko azalpena 4.1 atalean dago.
[68] Lematizazio erroreei buruz 5.2.1.6 atalean mintzatu gara.
[69] Katalogo honetara iristeko azalpena 4.1 atalean dago.
[70] Errore mota honi buruz 5.2.1.7 atalean mintzatu gara.

a. Erroreak diren agerpen horiek *Especial Case 7* bezala **etiketatuko** dira 3lb erabiliaz (5.2.1.7 ataleko irizpideak jarraituta).

b. Erroreak etiketatu ondoren, `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan dagokion fitxategian[71] (`landutakoak.n.ezusuak.txt` edo `landutakoak.n.usuenak.txt`), **`ERR`** marka jarriko zaio.

[71] Katalogo honetara iristeko azalpena 4.1 atalean dago.

A ERANSKINA: EuskalWordNet-en orrazketa: Editorearen eskuliburua

A.1 Sarrera

Donostiako Informatika Fakultateko Lengoaia Naturalaren Prozesamendurako IXA taldea, beste zenbait lanen artean, EuskalWordNet-en proiektua lantzen ari da. Lauzpabost urte daramatza lan honetan.

Esan beharra dago EuskalWordNet EuroWordNet abiapuntutzat hartuta sortu zela, ingeleseko kontzeptuei euskarakoak lotuz. Hasieran, euskararako WordNet egiteko, metodo erdiautomatikoak erabili ziren: *Morris Hiztegi*an agertzen ziren ingeleseko euskal ordainak sartu ziren. Horrela, ingeleseko synset-ei euskarako ordainak lotu zitzaizkien.

Ondoren, hurrengo pausoa honako hau izan zen: *Elhuyar Hiztegi Txiki*ko adierak EuskalWordNet-en agertzea. Ideia da hiztegi horretan euskal kontzeptu eta ordain nagusienak jasotzen direla eta horiek behintzat EuskalWordNet-en egon behar dutela.

Jarraian PROFIT proiektua sortu da, eta oraingoan helburua hauxe da: EuSemCor euskarako corpusaren eskuzko etiketatze semantikoa burutzea. 300.000 hitzez osaturiko corpusa da. Etiketatze semantikorako EuskalWordNet-eko synset erabiltzen dira, baina etiketatze-prozesuaz gain, helburua da corpusean agertuz joango diren adiera berri guztiak EuskalWordNet-ek jasotzea, eta horrela, EuskalWordNet **orraztea** eta **aberastea**.

Esan daiteke orain arte pertsona bakarra aritu dela editore lanetan (tarteka bigarren bat), baina lan honetan aritzeko lagun gehiagoren beharra ikusita, EuskalWordNet-en orrazketaren inguruko informazio guztia dokumentatzea erabaki da. Honenbestez, *Editorearen Eskuliburua* sortzea da lan honen helburua: EuskalWordNet ezagutzea, orrazketarako baliabideak erabiltzen jakitea, arazoei eta zalantzei aurre egiteko irizpideak edukitzea, ...

A.2 EuskalWordNet-en erabilera

A.2.1 Kokapena

Zer da EuskalWordNet? Bada, Euskarako Ezagutza Base Lexikal bat da (EBL). Bertan hitzei eta adierei buruzko informazioa jasotzen da eta hierarkikoki antolatuta daude. Antolamendua sinonimian oinarrituta dago: sinonimo multzo bakoitza, synset deritzona, hitzen adierez eratuta dago. Gainera, synset-en artean erlazio lexikal anitz daude, baina batez ere hiperonimia eta hiponimia dira landuta daudenak.

Sarreran esan dugun bezala, ingeleseko WordNet hartu zen abiapuntutzat, WordNet 1.5 hain zuzen ere. Duela gutxi arte hau izan da erabili den bertsioa, hau da, EuskalWordNet 1.5-a. Baina 1.5 bertsioaren ondotik ingeleseko 1.6 bertsioa kaleratu zen eta dagoeneko EuskalWordNet 1.6 bertsioa erabilgarri dago, bai kontsultarako, bai

orrazketarako (http://siuc02.si.ehu.es/wei/wei.html). Hau da une honetan erabiltzen den bertsioa.

Honetaz gain, aipatu beharra dago orain arte IZENak bakarrik daudela landuta; baina ADITZAK, ADJEKTIBOAK eta ADBERBIOAK ere landuko dira[72].

A.2.2 EuskalWordNet-en interfazea erabiltzeko argibideak

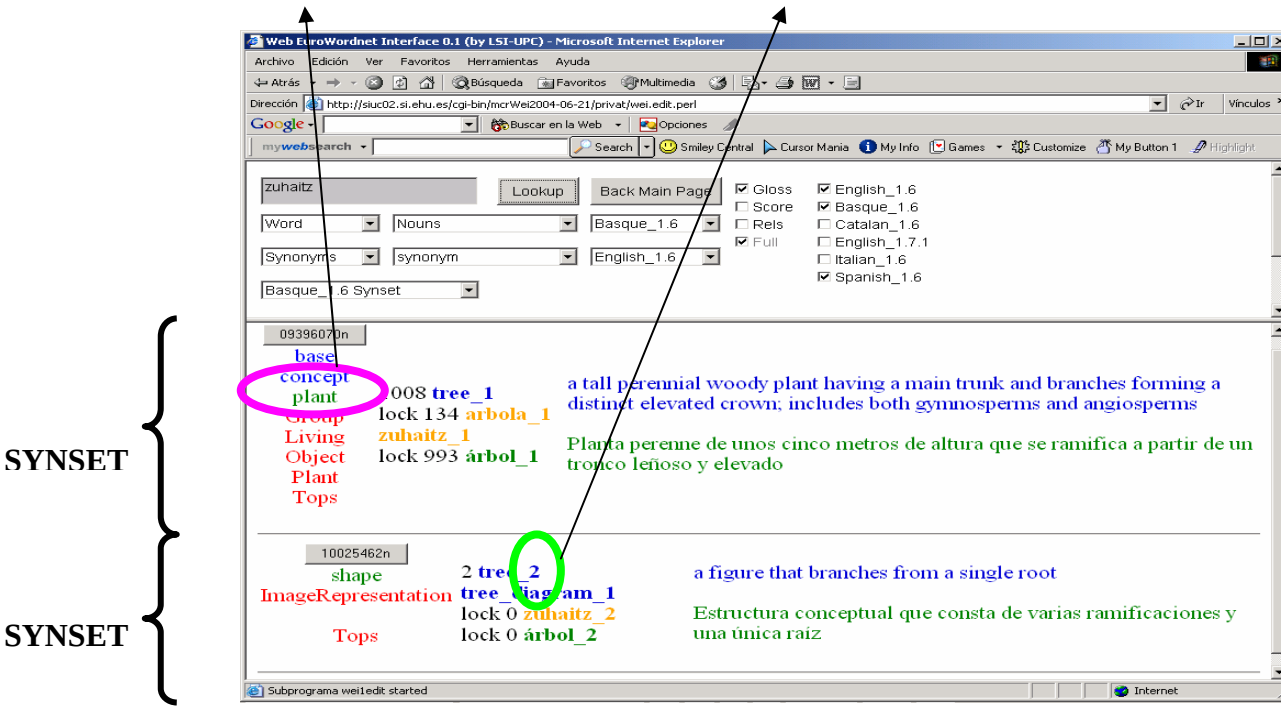
A.2.2.1 Oinarrizko kontzeptuak

SYNSET:

- Kontsultatu nahi dugun hitzaren adiera ezberdin bakoitzari **synset** bat dagokio, eta interfazean marra batez bereizirik agertzen da. Adibidean ikus daitekeen bezala, *zuhaitz* hitzak bi synset ditu, hau da, bi adiera: 'arbola' eta 'diagrama'.
- Bestalde, synset bakoitzak **synset-zenbaki** bat izango du.

VARIANT:

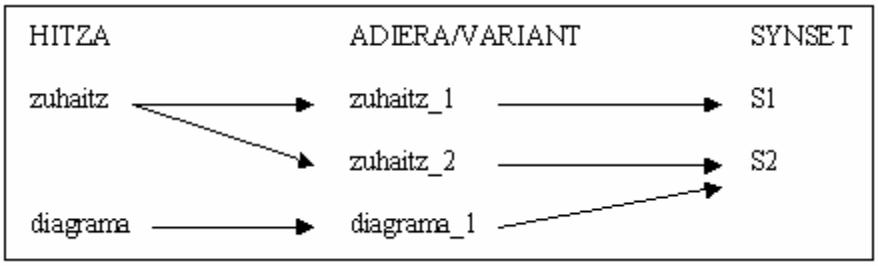
- Synset bakoitzean hizkuntza bakoitzeko dagoen ordaina.
- Ordain bakoitzak adiera-zenbaki bat du. Beheko irudian adibidez, lehenengo synset-ean, variant-ak hurrengoak dira: ingelesekoa, ***tree_1***, gaztelaniakoa ***árbol_1*** eta euskarakoak ***zuhaitz_1*** eta ***arbola_1***.



1go irudia

[72] Eskuliburu hau, izenen orrazketan oinarrituta dagoenez, irizpide asko izenen ediziorari bakarri degokie. Beste kategorien orrazketarekin hastean, irizpide berriak sortuko direla aurreikusten dugun arren, eskuliburu honetan agertzen diren irizpide asko eta asko kategoria guztientzat erabilgarriak izan daitezkeela susmatzen dugu.

Hurrengo eskeman ikusten da HITZA – ADIERA – SYNSET terminoen arteko erlazioa:



2. irudia

Ezkerretik eskuinera begiratuz gero (synset-en ikuspegitik), *zuhaitz_1*, *zuhaitz_2* eta *diagrama_1* VARIANT-ak lirateke. Alderantziz, hau da eskuinetik ezkerrera (hitzaren ikuspegitik) horiek ADIERAK lirateke.

Harreman semantiko nagusienak:

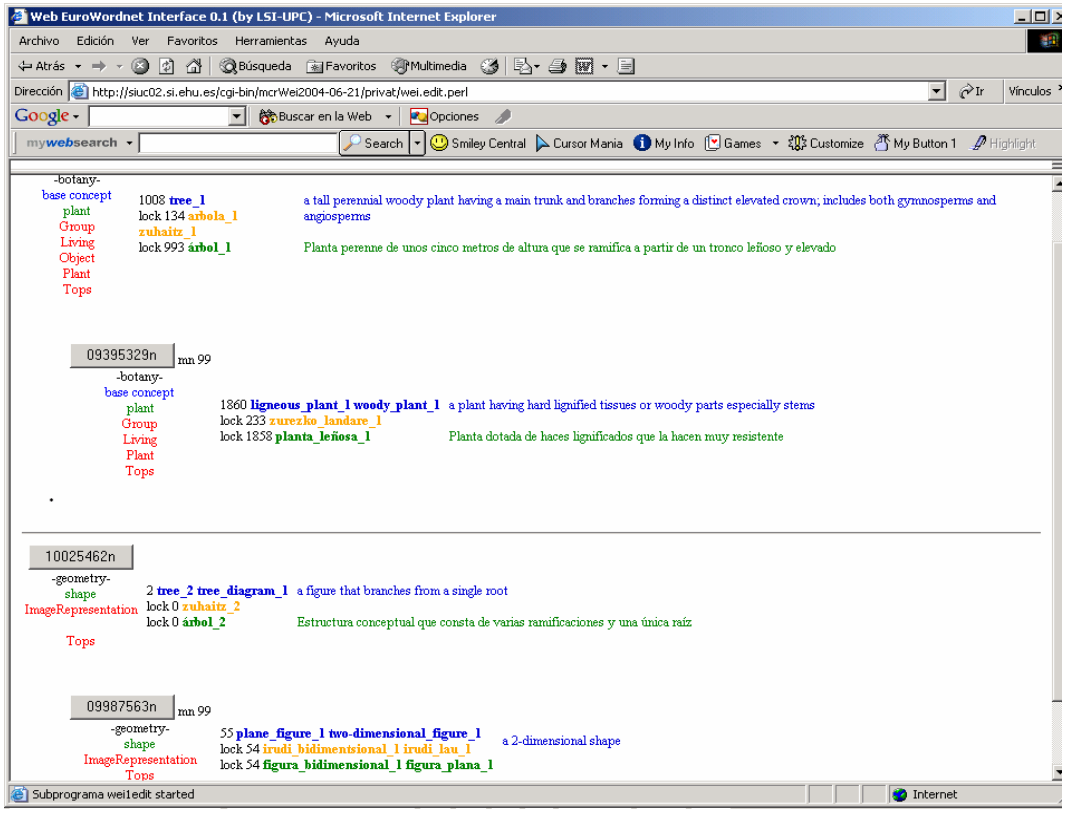
SINONIMIA:

- Hitz, synset edota variant baten sinonimoak, synset bakoitzean dauden variant-ak izango dira, eta ez agertzen diren synset-ak (hauek adiera ezberdinak baitira). Adibidez, *zuhaitz* hitzak bi adiera ezberdin ditu (bi synset), eta *zuhaitz* hitzaren sinonimoak, adiera horietako bakoitzean dauden euskal variant-ak izango dira. Esate baterako, *zuhaitz_1*-en sinonimoa *arbola_1* da, eta *zuhaitz_2*-k ez du sinonimorik (ikus goiko irudia).

HIPERONIMIA:

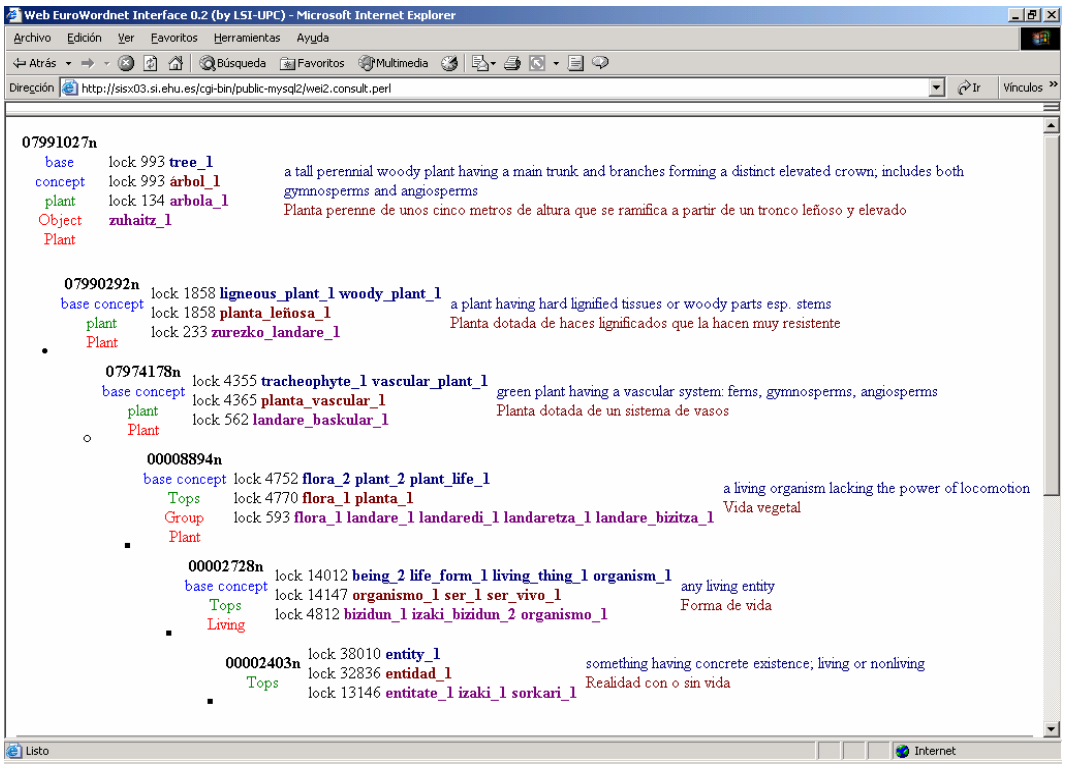
- Hitz, synset edota variant baten hiperonimoak eskatzen ditugunean, hauek baino orokorrago edo generikoagoak diren terminoak eskatzen ari gara. Adibidez: *zuhaitz* ➔ *zurezko landare* ➔ *landare*

- **Hiperonimo hurbilak vs Hiperonimo kate osoa**: Hurrengo adibidean ikus daitekeen bezala, *zuhaitz_1*-en hiperonimo hurbilena *zurezko_landare_1* izango litzateke eta *zuhaitz_2*-rena *irudi_bidimentsional_1* edo *irudi_lau_1*.



3. irudia

Eta *zuhaitz_1*-en hiperonimo-kate osoa eskatuz gero, terminoen arteko harremana **zehatzenetik orokorrenera** eskatuko genukeen, hots, *zuhaitz_1*-en hiperonimo guztiak:

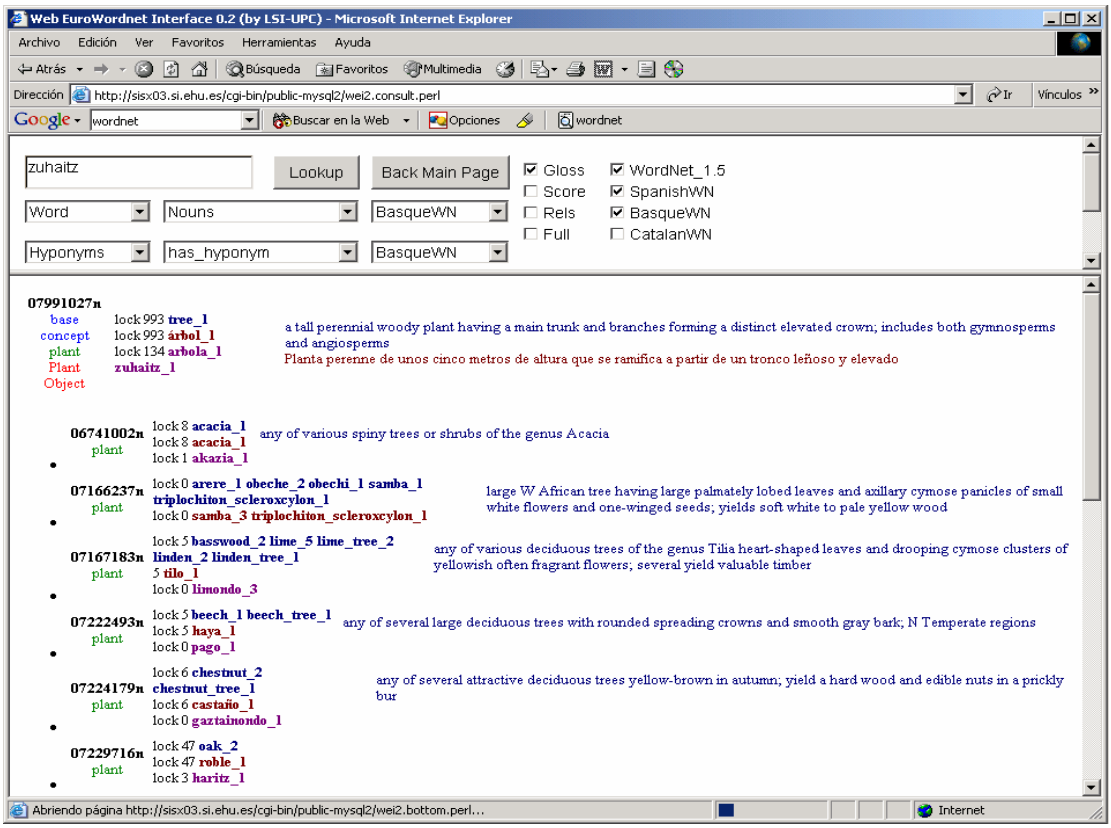


4. irudia

HIPONIMIA:

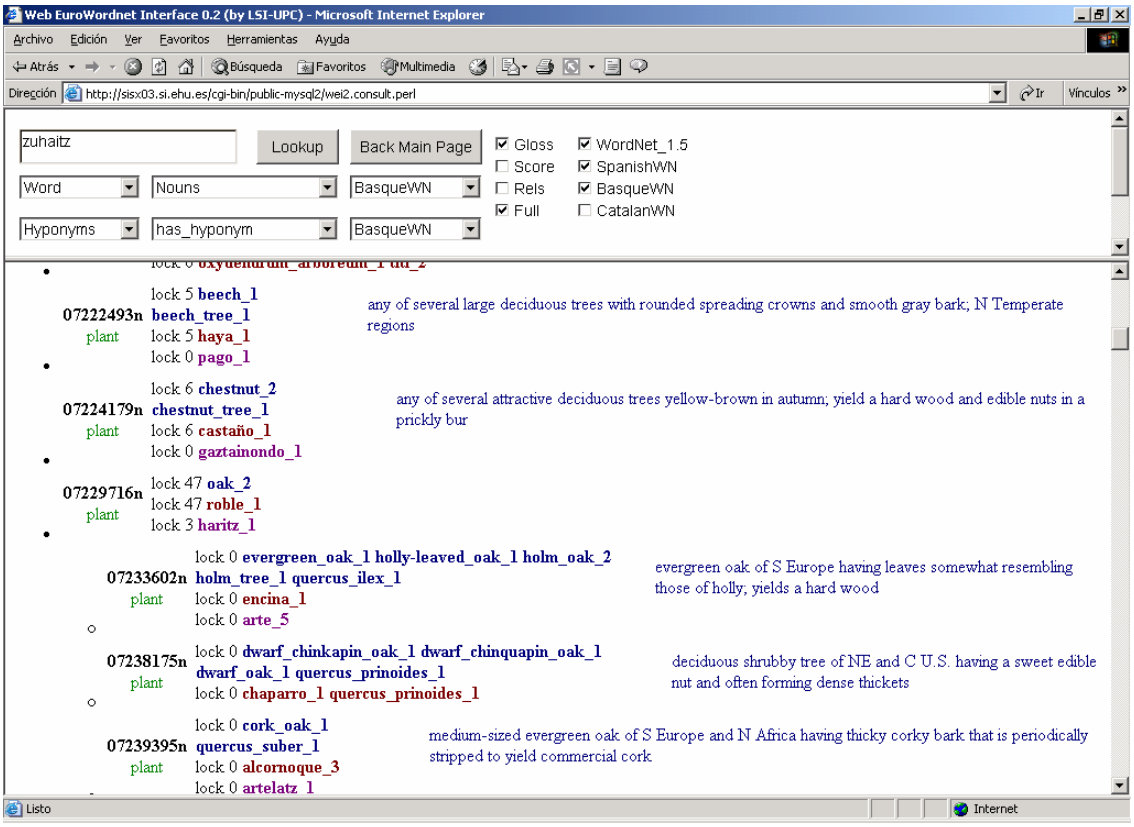
- Hitz, synset edota variant baten hiponimoak eskatzen ditugunean, termino orokor honek bere baitan hartzen dituen termino espezifikoak eskatzen gara. Adibidez, *zuhaitz*en hiponimoak 'zuhaitz motak' izango dira: *zuhaitz* ➔ *pago*

- **Hiponimo hurbilak vs Hiponimo zuhaitz osoa:** Adibidean ikus daitekeen bezala, *zuhaitz_1*-en hiponimo hurbilak, *akazia_1*, *limondo_3*, *pago_1*, *gaztainondo_1*, *haritz_1*, eta abar dira.



5. irudia

Eta *zuhaitz_1*-en hiponimo-zuhaitz osoa eskatuz gero, terminoen arteko harreman osoa, **orokorrenetik zehatzenera**, 6. irudikoa litzateke. Bertan ikus daitekeen bezala, *zuhaitz* mota ezberdinak daude (hiponimo hurbilak): *pagoa*, *gaztainondoa*, *haritza* eta abar. Eta hiponimo hurbil hauek, aldi berean, beste hiponimo batzuk izan ditzakete, esate baterako, *pago* mota ezberdinak egon daitezke: *artea*, *artelatza*, eta abar. Hala, hiponimo-zuhaitz osoa eskatuz gero, synset baten hiponimo hurbilak ikus ditzakegu, hiponimo hurbil hauen hiponimoekin batera.

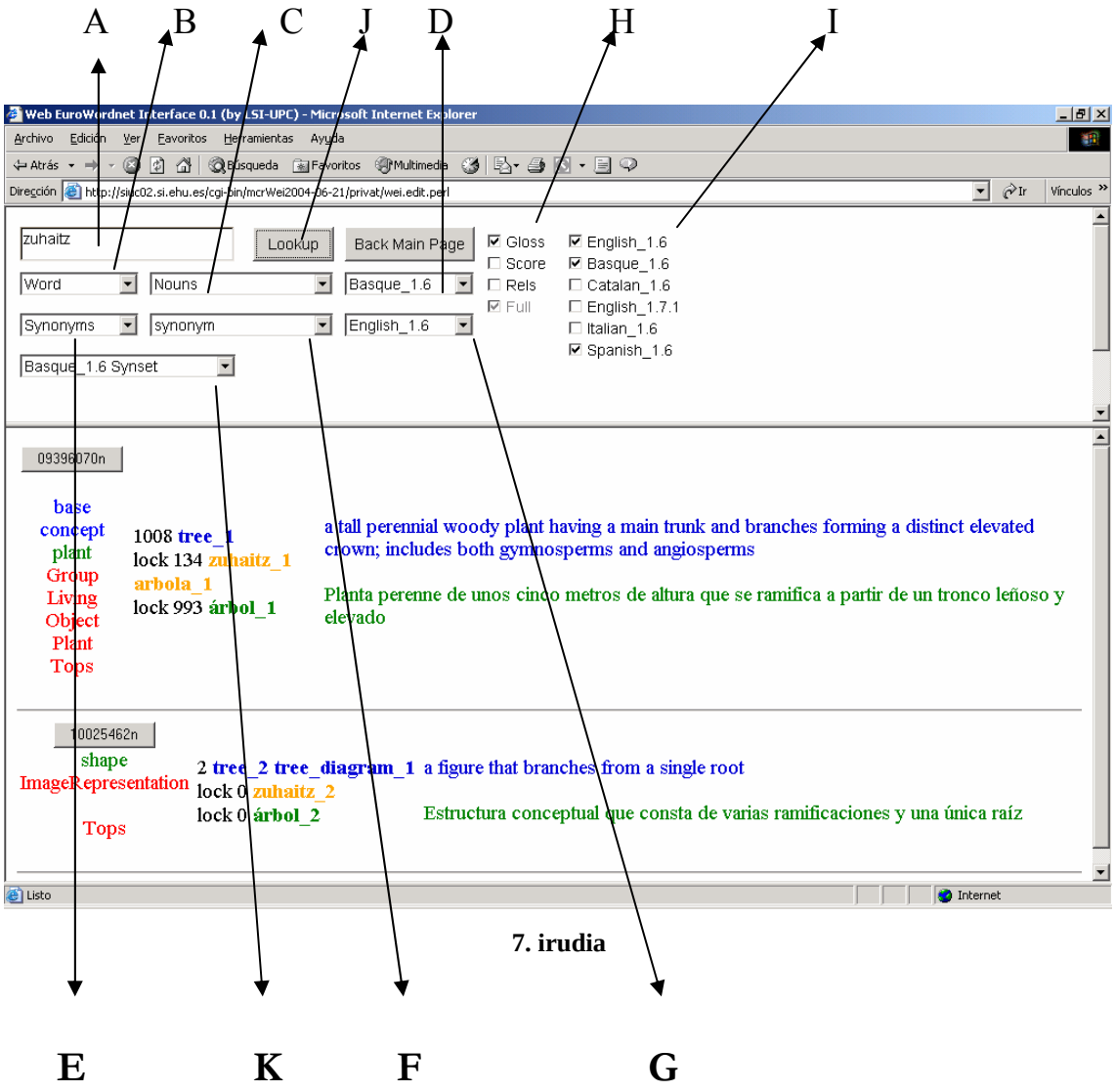


6. irudia

A.2.2.2 Nola egin bilaketa

Ondoren, EuskalWordNet-en kontsultak egiteko argibideak ematen dira, hau da, bilaketak nola egin eta informazio mota desberdinak nola lortu.

Hurrengo irudian interfazearen funtzio garrantzitsuenen azalpenak ematen dira:



7. irudia

A = Bilaketarako testu-kutxa.

B = A testu-kutxan idatzitako kontsulta, hitza, synset-a edo variant-a den zehazten da:

Word (*zuhaitz*), **Synset** (*07991027*) edo **Variant** (*zuhaitz_2*).

C = **A** testu-kutxan idatzitakoaren kategoria zehazteko balio du:

Noun / Verb / Adjective / Adverb

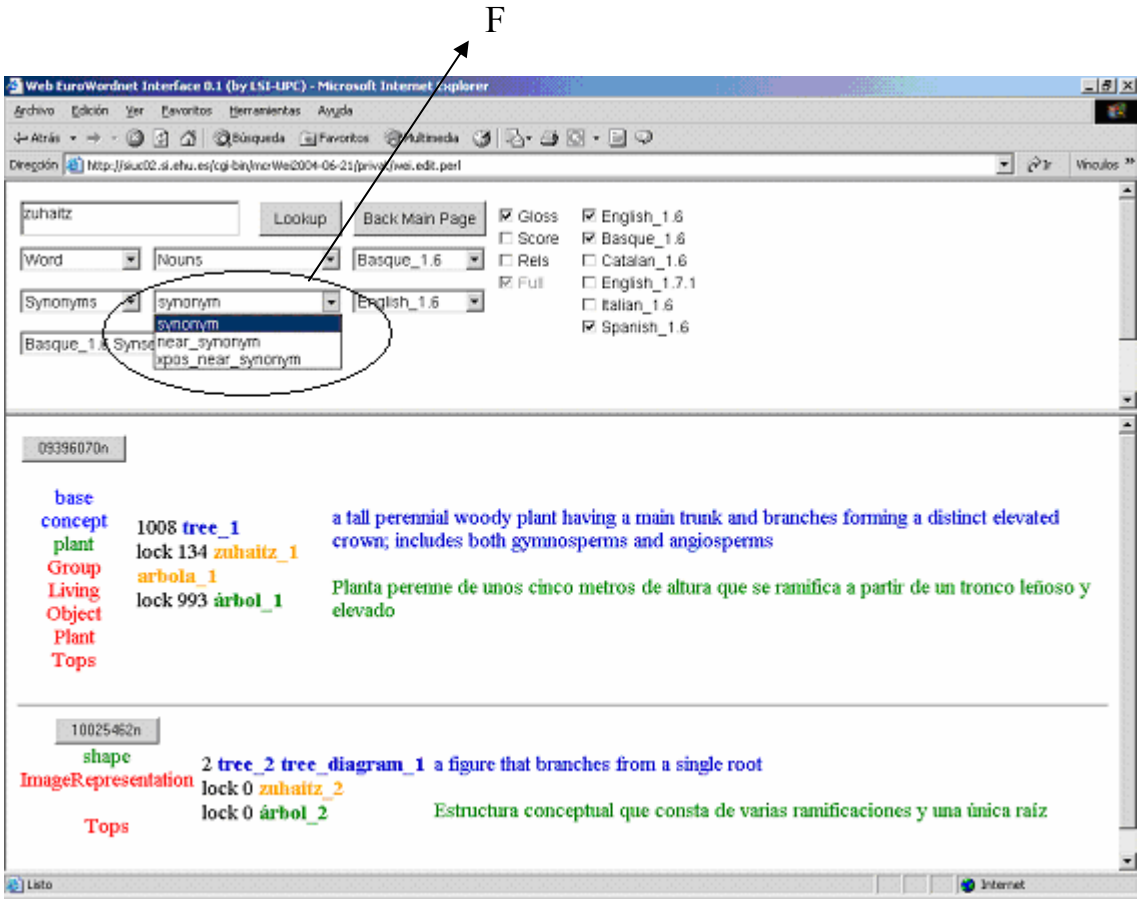
D = **A** testu-kutxan idatzitakoa zein WordNet-en bilatu nahi dugun adierazten du:

English_1.6 / Basque_1.6 / Catalan_1.6 /
English_1.7.1 / Italian_1.6 / Spanish_1.6

E = **A** testu-kutxan idatzitako hitz, synset edo variant horrekiko harreman mota desberdinak bilatzeko aukera ematen du:

hyponyms, hyperonyms, meronyms, antonyms, holonyms, e.a.

F = Nahiz eta **E**-n aukeratutakoaren arabera **F** automatikoki aldatu egiten den, **F**-k **E**-ren zehaztapenerako aukera batzuk ematen ditu, adibidez:



8. irudia

G = Zehaztutako harreman semantikoa zein WordNet-en ikusi nahi den adierazten du. Hemen kontuan izan beharrekoa da, EuskalWordNet ingeleseko WordNet-an oinarrituta garatzen ari dela, eta kontzeptu batzuk euskaratu gabe egon daitezkeela. Arrazoi horregatik, eremu honetan English_1.6 (ingeleseko WordNet-a, alegia) jartzea gomendatzen da. Horrela, ezagutza-basean dauden synset guztiak ikusiko ditugula ziurtatzen dugu.

H = Kontrol-lauki hauen bidez, pantailan informazio gehiago edo gutxiago ikusteko aukera ematen du.

- **Gloss**: Synset-aren adibide edo definizio laburra ikusteko aukera ematen du.

- **Score**: Konfiantza-nerria ikusteko aukera ematen du[73].

- **Rels**: Synset-ak izan ditzakeen harreman semantiko mota guztiak ikusteko aukera ematen du[74].

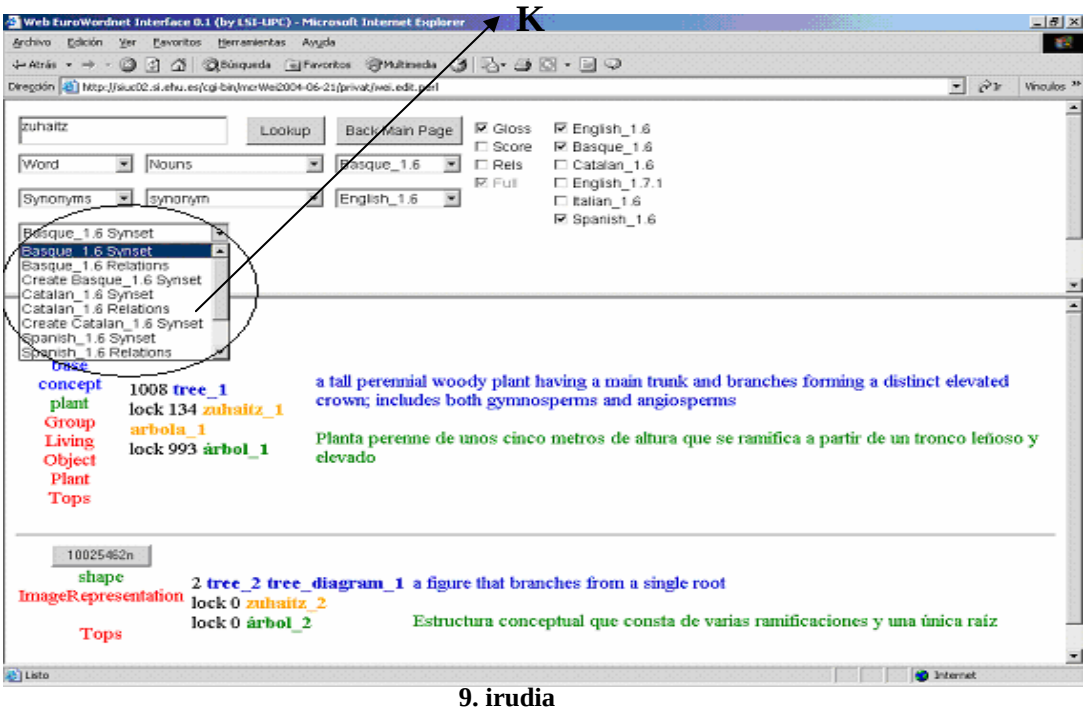
- **Full**: Honi sakatuta, synset-en harreman semantikoen agerpena era ezberdinetara eska daiteke:
    - beraien osotasunean (hiperonimo kate / hiponimo zuhaitz osoa, adibidez)
    - harreman hurbilenak bakarrik (hiperonimo / hiponimo zuzenak)

I = Hauen bitartez kontsultaren emaitza zein hizkuntzatan ikusi nahi dugun erabaki dezakegu: **English_1.6** edota **Basque_1.6** edota **Catalan_1.6** edota **English_1.7.1** edota **Italian_1.6** edota **Spanish_1.6**

J = Behin hautaketa eginda, botoi hau sakatu behar da bilaketari hasiera emateko.

K = Hizkuntza desberdinetan synsetak aldatzeko, sortzeko eta informazioa jasotzeko aukerak ematen dituzte. Editorearentzat **Basque** aukerak izango dira erabilgarrienak; hiru dira:

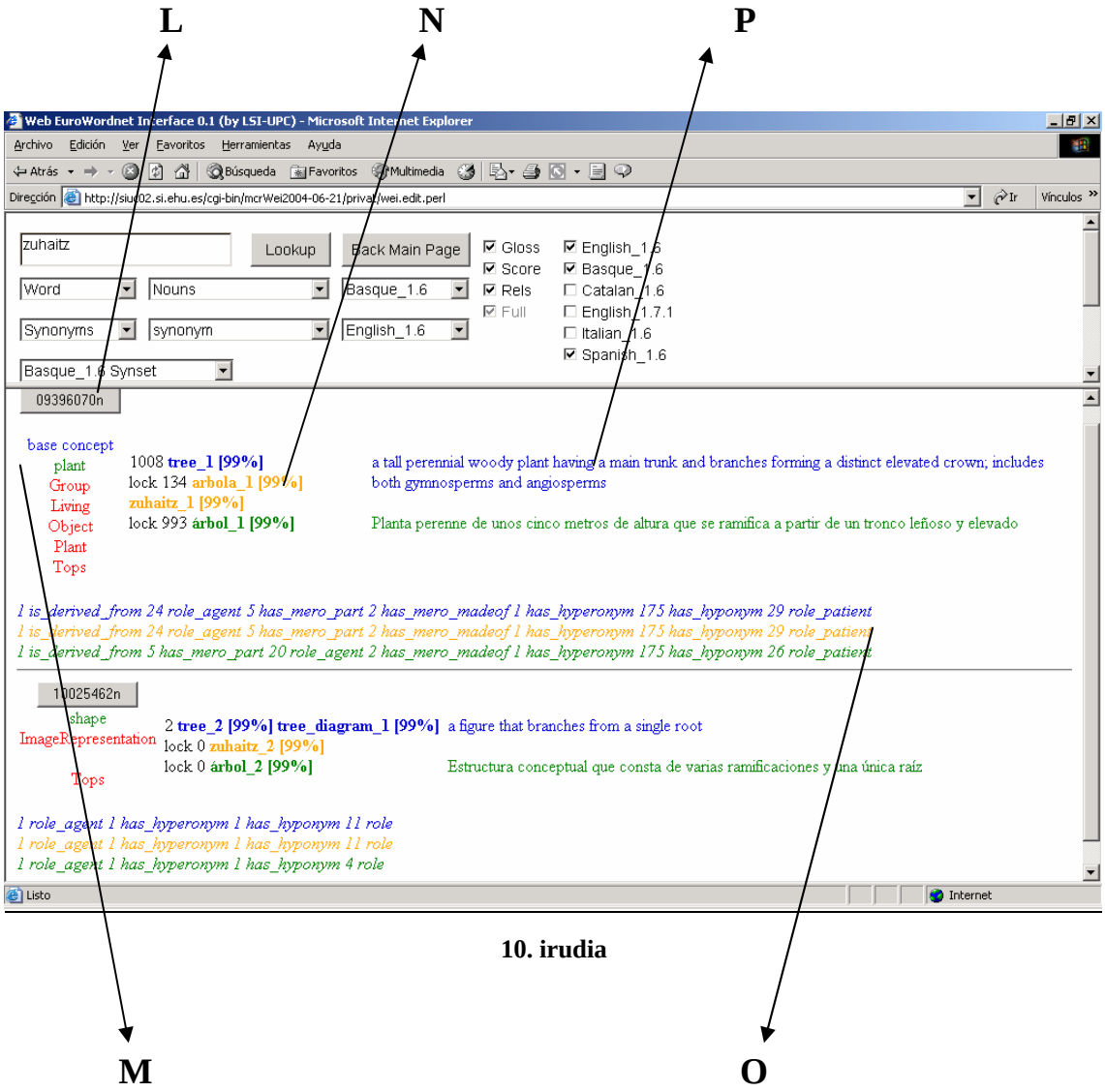
- **Basque_1.6 Synset**: Euskal synsetetan aldaketak egin daitezke: variant berriak gehitu, edo variant-ak ezabatu, ...
- **Basque_1.6 Relations**: synset-ek besteekiko dituzten erlazio semantikoen berri ematen du.
- **Create Basque_1.6 Synset**: euskaraz synset berri bat sortzeko balio du.



9. irudia

[73] Konfiantza-neurriari buruzko azalpena A.2.2.3 atalean dator.
[74] Harreman semantikoei buruzko azalpena A.2.2.3 atalean dator.

A.2.2.3 Nola interpretatu bilaketaren emaitza



10. irudia

L = Synset-zenbakia.

M = Synset-aren eremu semantikoak. Hiru motatako eremuak ager daitezke:

Base Concept: oinarrizko kontzeptua denean agertuko da bakarrik; beti urdinez. Banaketa semantiko sinplea: sailkapen semantiko mota bat; beti berdez: artifact, plant, shape,...

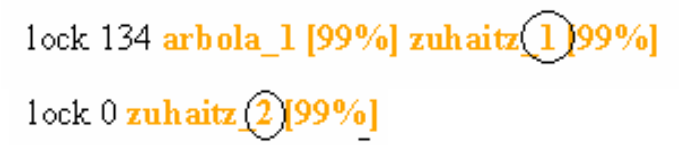
Banaketa semantiko aberatsa edo Top Ontology: sailkapen semantiko aberatsagoa; beti gorriz: Artifact, Plant, Object, ...

N = Synset horri dagozkion variant multzoa, **I**-n egindako aukeren arabera: ingelerazkoak **urdinez**; gaztelaniakoak **berdez**; euskarakoak **laranjaz**; katalanerazkoak **gorriz**; italierazkoak grisez.

- **Lock**: eskuz landua izan dela adierazteko marka[75].

- **Lock-en ondoan dagoen zenbakia**: hizkuntza horretako synset-ak dituen hiponimo-kopurua adierazten du; adibidez, *zuhaitz*ek, 'landare' adierarekin 134 hiponimo ditu:

- **Adiera-zenbakia**: hitzaren adiera ezberdinak zenbakien bidez desberdintzen dira. *Zuhaitz*ek bi adiera ditu, 'landarearena' eta 'diagramarena'; beraz, adiera-zenbaki desberdina beharko dute:



- **% 99**: konfidantza-neurria. Eskuz landu direnak eman daitekeen ehunekorik altuena izango dute: %99, alegia.

O = Hizkuntza bakoitzeko WordNet-eko synset-ek dituzten harreman semantikoen kopuruak erakusten ditu. Esate baterako, *zuhaitz_1*-ek EuskalWordNet-en honako harreman semantikoak ditu:

1 is_derived_from 24 role_agent 5 has_mero_part 2 has_mero_madeof
1 has_hyperonym 175 has_hyponym 29 role_patient

P = Synset-aren azalpen laburra, bere adiera ulertzeko baliagarria dena.

[75] Interfaze publikoan, *Lock* dauden synset-ak bakarrik ikus daitezke. *Unlock* edo landugabe daudenak, interfaze pribatuan bakarrik daude atzigarri.

A.3 Editore-lana

A.3.1 Baliabideak

Editoreak zenbait baliabide izango ditu EuskalWordNet-en orrazketarako, eta atal honetan zerrendatuko ditugu:

A.3.1.1 EuskalWordNet

Lan honen hasieran esan bezala, gaur egun EuskalWordNet 1.6 bertsioan egiten da lan: http://siuc02.si.ehu.es/wei2004-06-21/wei.html

Askotan oso baliagarria izango zaio editoreari Browser-eko ingeleseko Wordnet 1.6-ra jotzea, EuskalWordNet 1.6-n dagoen informazioa beste honetan kontrastatzeko. Gainera, adibideak bertan aurki daitezke; eta bilaketak egiteko askoz azkarragoa dela ere esan daiteke. Ingeleseko WordNet 1.6 Browser-a exekutatzeko `sisx01`[76] makinan idatzi:

```
1-[sisx01 ~]% wn16
2-[sisx01 ~]% wnb &
```

Esan beharra dago, ingelesez WordNet 2.0 bertsioan ari direla lanean dagoeneko. Editorea bertsio horretara jo dezake 1.6 bertsioan aurkitzen ez duen zerbait kontsultatzeko, ze batetik bestera aldaketak egon baitaitezke.

WordNet 2.0: http://www.cogsci.princeton.edu/cgi-bin/webwn

Edo `sisx01` makinan agindu honen bidez ere exekutatu daiteke:

```
3-[sisx01 ~]% wn20
4-[sisx01 ~]% wnb &
```

A.3.1.2 Euskarako hiztegiak

Hauek elebakarrak eta elebidunak izan daitezke:

- *Elhuyar Hiztegi Txikia* (paperean)
- *Elhuyar Hiztegia (elebiduna)*
- *Elhuyar Hiztegi Entziklopedikoa* (http://www1.euskadi.net/hizt_el/indice_c.htm)
- *Euskal Hiztegia*
- *Euskaltzaindia* (http://www.erabili.com/lantresnak/hiztegiak/euskaltzaindia)
- *Euskalterm Hiztegi Terminologikoa* (http://www1.euskadi.net/euskalterm/indice_c.htm)

[76] `Sisx01` makina erabiltzeko *shell*-a behar da (ikus 3. atala).

A.3.1.3 EDBL Datu-base lexikala

http://ixa2.si.ehu.es/edbl/

A.3.1.4 Gaztelaniako hiztegiak

- *Diccionario de la Lengua Española* (http://www.rae.es/)

A.3.1.5 Ingeleseko hiztegiak

Hauek elebakarrak eta elebidunak izan daitezke:

- *Collins* (paperean)
- *Oxford* (paperean)
- *Wordreference* (http://www.wordreference.com/)
- *Cambridge* (http://dictionary.cambridge.org)
- *Morris Hiztegia* (http://www.hiztegia.net/)
- *Onelook* (http://www.onelook.com/)

A.3.1.6 Corpusak

- *XX. mendeko Euskararen Corpus Estatistikoa* (http://www.euskaracorpusa.net/)
- *Ereduzko prosa gaur* (http://www.erabili.com/lantresnak/aztergailuak/prosa)
- *EuSemCor euskara corpusa* (3lb tresna edota ./`kwic` agindua erabilita[77])

A.3.1.7 Hiztegixa

Hiztegixa IXA Taldeak sortutako tresna bat da; bertan hiztegirik garrantzitsuenak eta erabilgarrienak jasotzen dira. Beraz, arestian aipatutako zenbat hiztegi jasotzen ditu. Exekutatzeko ondorengo hau da agindua `sisx01` makinan[78]:

```
5-[sisx01 ~]% hiztegixa
```

```
****** M O R R I S ******          | '1': Eusk -> Ing   | '5': Euskal Hiztegia |
 *                           *      | '2': Ing  -> Eusk  | '6': Sinonimoak      |
 *  EUSKARA -> INGELESA      *      | '3': Eusk -> Gazt  | '7': Euskaltzaindia  |
 *                           *      | '4': Gazt -> Eusk  | '8': Lematizatzailea |
 *                           *      | 'A': Modernoa      | '9': Corpus          |
 *************************          | '.': Bukatzeko     | '10': Guztiak!!!<--  |
```

Ezkerretara, une horretan martxan dagoen hiztegia; kasu honetan lehenengoa, *Morris*. Eskuinetara gainontzeko guztiak.

---

[77] Ikus 3. atala.

[78] `Sisx01` makina erabiltzeko *shell*-a behar da (ikus 3. atala).

Hiztegixa IXAko web orrian (pribatuan) ere eskuragarri dago: http://ixa2.si.ehu.es/hiztegixa/

A.3.2 Hitz baten orrazketarako prozesua

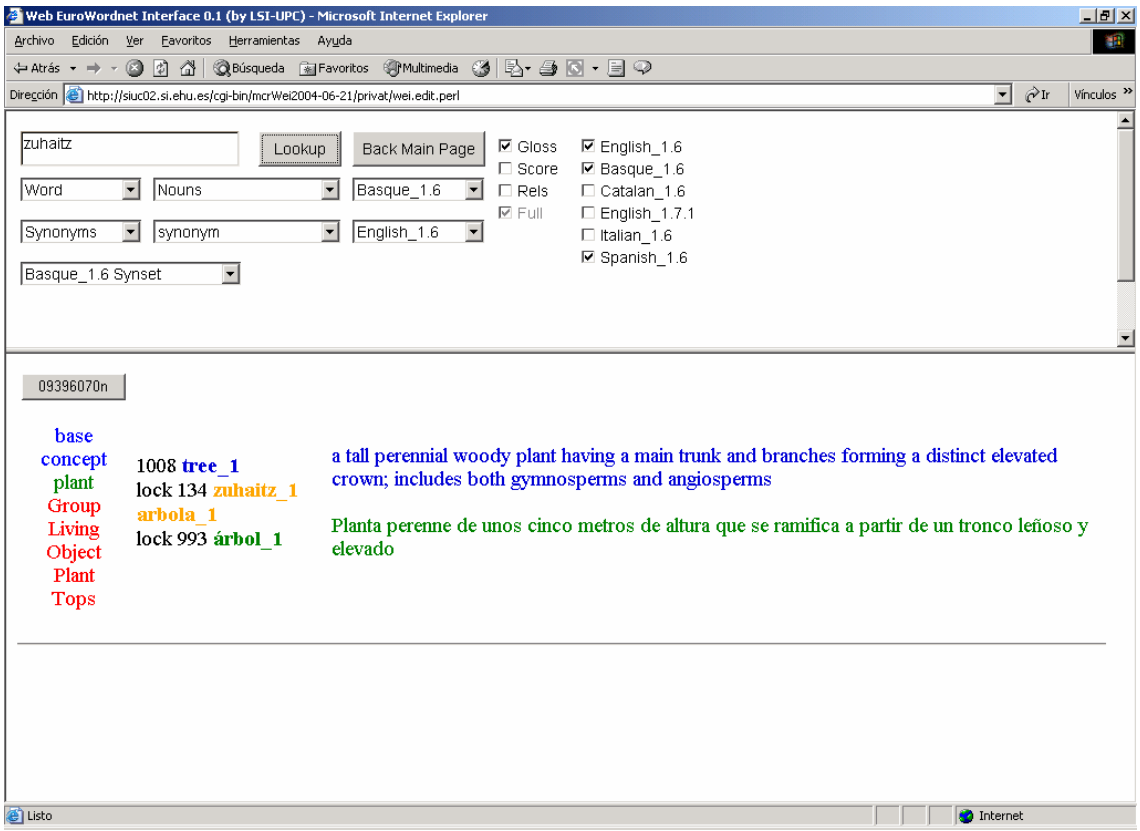
Sarreran aipatu bezala, orain arte izenak dira EuskalWordNet-en landuta daudenak. Izen eta adiera horiek gutxienez, *Elhuyar Hiztegi Txiki*ak jasotzen dituenak dira, arruntenak horiek direla uste baita. Ondoren, adjektibo**,** adberbio eta aditzen lanketari ekingo zaio.

Atal honetan hitz baten orrazketan eta lanketan editoreak jarraitu behar dituen pausoen azalpena dator. Adibide gisa erabiliko dugun hitza *zuhaitz* izango da.

Lehenengo eta behin, editoreak hitz hori EuskalWordNet-en landuta dagoen ala ez jakin behar du. Horretarako bertara joko du eta *zuhaitz* hitzaren bilaketa egingo du. Bi gauza gerta daitezke:

- EuskalWordNet-en egotea
- EuskalWordNet-en ez egotea

Demagun, lehenengo kasuan bezala *zuhaitz* hitza landuta dagoela, eta EuskalWordNet-en honela ageri dela:



11. irudia

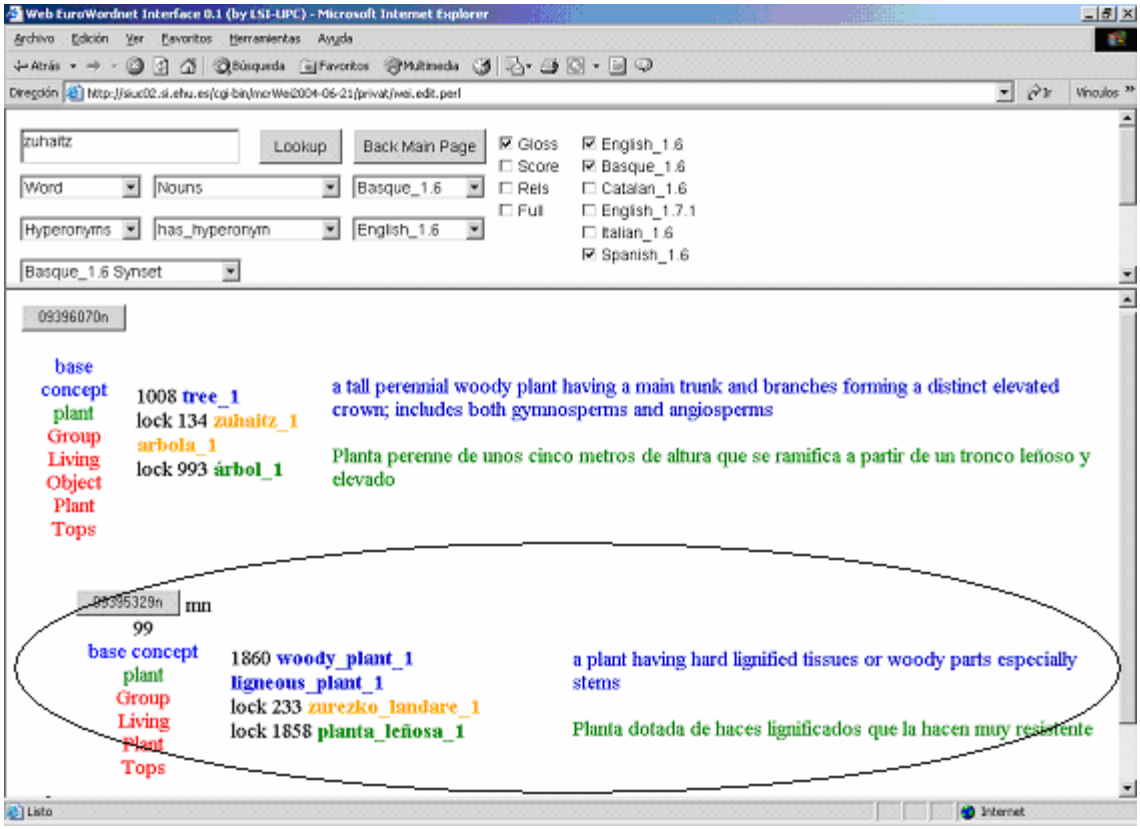
A.3.2.1 *Synset-en ulermena*

Lehenengo pausoa agertzen diren synset-ak ulertzea da. Ikus daitekeenez, EuskalWordNet-en *zuhaitz* hitzak synset bakarra du, 'landarea' adierazten duena:

09396070n zuhaitz_1, arbola_1 = 'landarea'

Kasu honetan ulerterraza gertatzen da *zuhaitz* hitzaren synset-a. Gerta daiteke, ordea, batzuetan mota desberdinetako zailtasunak sortzea: synset ilunak, zenbait synset-en artean bereizketarik ez ikustea, hiperonimo eta hiponimoetan hitz bera agertzea, besteak beste. Honelako kasuak aurrerago landuko dira (4. atalean), hartutako erabakiak eta irizpide nagusiak banan-banan azalduz.

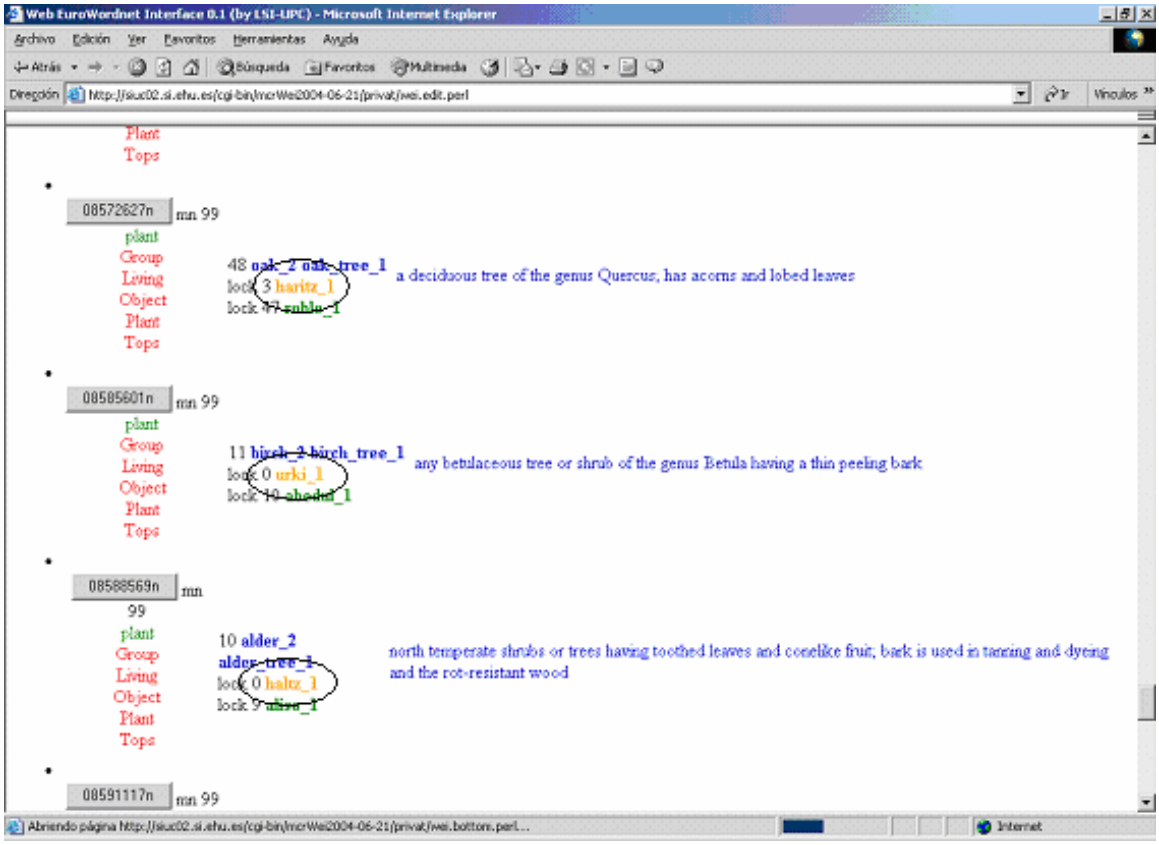
Hala eta guztiz ere, oso lagungarria izaten da bilaketan bere hiperonimorik hurbilena, edota hiperonimo-kate osoa jasotzea. Honela:



12. irudia

Irudi honetan, *zuhaitz_1*-en hiperonimorik hurbilena ikus daiteke; eta adiera hobeto ulertzen lagun dezake: *zurezko_landare_1*.

Beste aukera bat izaten da ulertu nahi dugun hitzaren hiponimoak ikustea, adiera ulertzen laguntzeko.



13. irudia

Beraz, osatu dugu lehenengo urratsa: adieren ulermena.

A.3.2.2 Synset-en egokitasuna

Editoreak EuskalWordNet-eko adierak ulertu dituenean, beren egokitasuna aztertu behar du.

A.3.2.2.1 Euskarako hiztegi elebakar eta elebidunetara jo

Lehenik, editoreak euskarako hiztegietara joko du *zuhaitz* hitzak dituen adierak aztertzeko. Adibidez, *Elhuyar Hiztegi Txiki*an begiratuz gero, (arestian esan bezala bertan dauden izen eta adierak gutxienez agertu behar dute EuskalWordNet-en) honako emaitza hau agertzen da:

Elhuyar Hiztegi Txikia: *zuhaitz*

1. Árbol. "*Zuhaitz ugariz jantziriko lurraldea*".
2. (egitura, eskema) Árbol. "*Zuhaitz genealogikoa*".

Ikus daiteke, gure adibideak bi adiera dituela *Elhuyar Hiztegi Txiki*an. Lehenengoak 'landareari' egiten dio erreferentzia. Beraz, hau da EuskalWordNet-ek jasotzen duen adiera. Bigarrena, berriz, 'eskema' edota 'egitura' adierazteko balio

duena da. Eta hau ez du EuskalWordNet-ek jasotzen. Orduan, egokitasuna aztertzen denean, bi puntu lantzen dira:

- EusWordNet-en dauden synset-ak ea **egokiak** diren; eta *zuhaitz_1* halaxe gertatzen da, *Elhuyar Hiztegi Txiki*ko 1go adierarekin bat baitator.

- Adiera edota synset-en bat **faltan** edo **soberan** dagoen; eta kasu honetan, *Elhuyar Hiztegi Txiki*ko bigarren adiera falta da EuskalWordNet-en ('diagrama').

Hor ditugu baliabideen atalean aipaturiko beste hainbat tresna, *zuhaitz* hitzak dituen adierak egiaztatzeko: *Euskal Hiztegia*, *Euskalterm*, EDBL, besteak beste. Beraz, euskarako hiztegiak kontsultatu ondoren, baiezta daiteke *zuhaitz* hitzak bi adiera dituela. Eta, aurreko atalean ikusi ahal izan dugun bezala, EuskalWordNet-en *zuhaitz*en adiera bat agertzen da, baina bestea ez. Orduan, editoreak eman behar duen hurrengo pausoa hauxe da: adiera hori EuskalWordNet-en sartzeko **synset egokia** aurkitu, eta bertan txertatu. Ondorengo atalean datoz horretarako argibideak.

A.3.2.2.2 *Nola sartu euskal ordaina synset batean*

Lehendabizi, hiztegi elebidunetara jo behar da *zuhaitz* hitzaren itzulpena jasotzera: *Morris Hiztegia*, *Oxford* , *Elhuyar*, ... (ikus 3.1. atalean)

Bilaketa egin ondoren, *zuhaitz*en itzulpenak ditugu: *tree* eta *árbol*. Orduan, ingeleseko *tree* eta gaztelaniako *árbol* aztertu behar dira, euskarako eta beste hizkuntzetako kontzeptuek gauza bera adierazten dutela ziurtatzeko. Horretarako, ingeles eta gaztelaniako hiztegi elebakarretan begiratu behar da, hitz hauen adiera desberdinen definizioak euskarako definizioekin parekatzeko. Esate baterako *Euskal Hiztegi Moderno*ak *zuhaitz* horrela definitzen du:

- Zurezko landare bizikorra, altuera aldakorrekoa, baina sarritan handia. Zurtoina (enborra) lurretik urruti samar adarkatzen da eta espezie bakoitzaren bereizgarri den adaburua eratzen du"

- Elkarrekiko erlazionaturik dauden edo sistema bat osatzen duten hainbat elementuren arteko mailaz mailako hierarkia-erlazioa grafikoki adierazten duen egitura adarkatua (bereziki hizkuntzalaritzan eta informatikan erabiltzen da)

Wordreference-ko definizioak *tree*-rentzat hurrengoak dira:

- Any large woody perennial plant with a distinct trunk giving rise to branches or leaves at some distance from the ground

- A branching diagrammatic representation of something, such as the grammatical structure of a sentence

Eta azkenik, *árbol*-en definizioak *Diccionario de la Lengua Español*-ean ondorengoak dira:

- Planta perenne, de tronco leñoso y elevado, que se ramifica a cierta altura del suelo.
- Cuadro descriptivo, la mayoría de las veces en forma de árbol.

Adibide honetan, hizkuntza guztietako ordainen adierak bateragarriak dira, hau da *zuhaitz*ek eta honen itzulpenak diren *tree* eta *árbol*-ek, berdinak diren bi adiera dituzte. Hortaz, bi adiera horiek dituzten erdal ordain horien (*árbol* eta *tree*) synset-etan euskarako *zuhaitz* hitza txerta daiteke.

Hala ere, *zuhaitz* hitzaren kasuan, bi adiera adierazteko ordain bakarra dago bai euskaraz, bai gaztelaniaz eta bai ingelesez. Baina, badira kasu konplexuagoak, non euskal hitz batek adiera bat baino gehiago dituen, eta hitz eta adiera hauen itzulpenak ingelesez eta gaztelaniaz, ordain bat baino gehiago diren. Adibidez:

lur: (ingelesez)

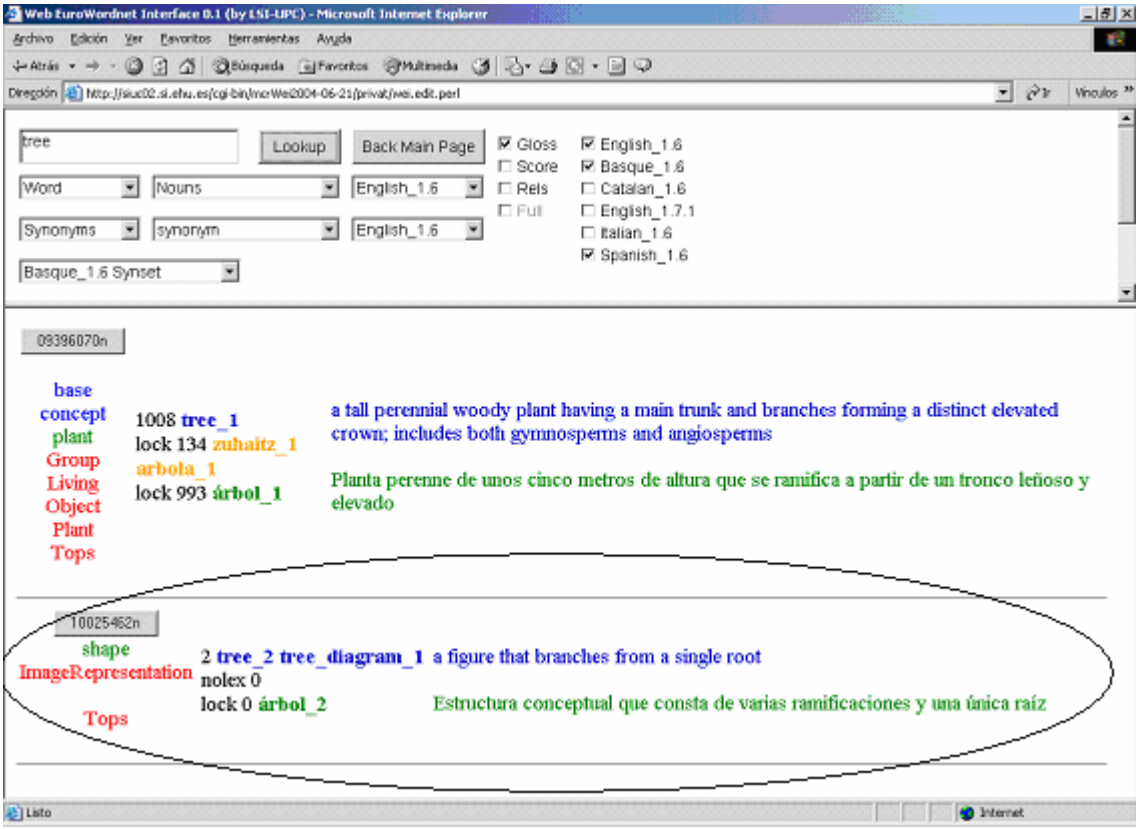
1. (Astron.) Earth.
2. (Kristau.) earth, world.
3. (ez airean) land.
4. (behekoa) ground.
5. (etxe barrukoa) floor.

lur: (gaztelaniaz)

1. tierra
2. suelo, tierra

Hauek izango lirateke formarik arruntenak. Eta hauek denak hiztegi elebakarretan aztertu behar dira, erdal ordain horien synset-etan euskarako *lur* hitza erabilgarria izan daitekeela egiaztatzeko.

Behin lantzen ari garen hitzaren (demagun, *zuhaitz*) eta dagozkion erdal ordainak (*árbol* eta *tree*) ezagututa, erdal ordain hauen synset-ak hizkuntza hauetako WordNet-en kontsultu behar dira, euskarako hitzari falta zaizkion adierak txertatzeko; *zuhaitz* hitzaren kasuan, adiera berri bat sartu behar da (‘diagrama’ri dagokiona, hain zuzen ere). Horretarako, esan dugun bezala, *tree* edo *árbol* hitzak bilatu behar dira ingeleseko eta gaztalaniakoWordNet-etan. Bai batean, bai bestean hau da emaitza:



14. irudia

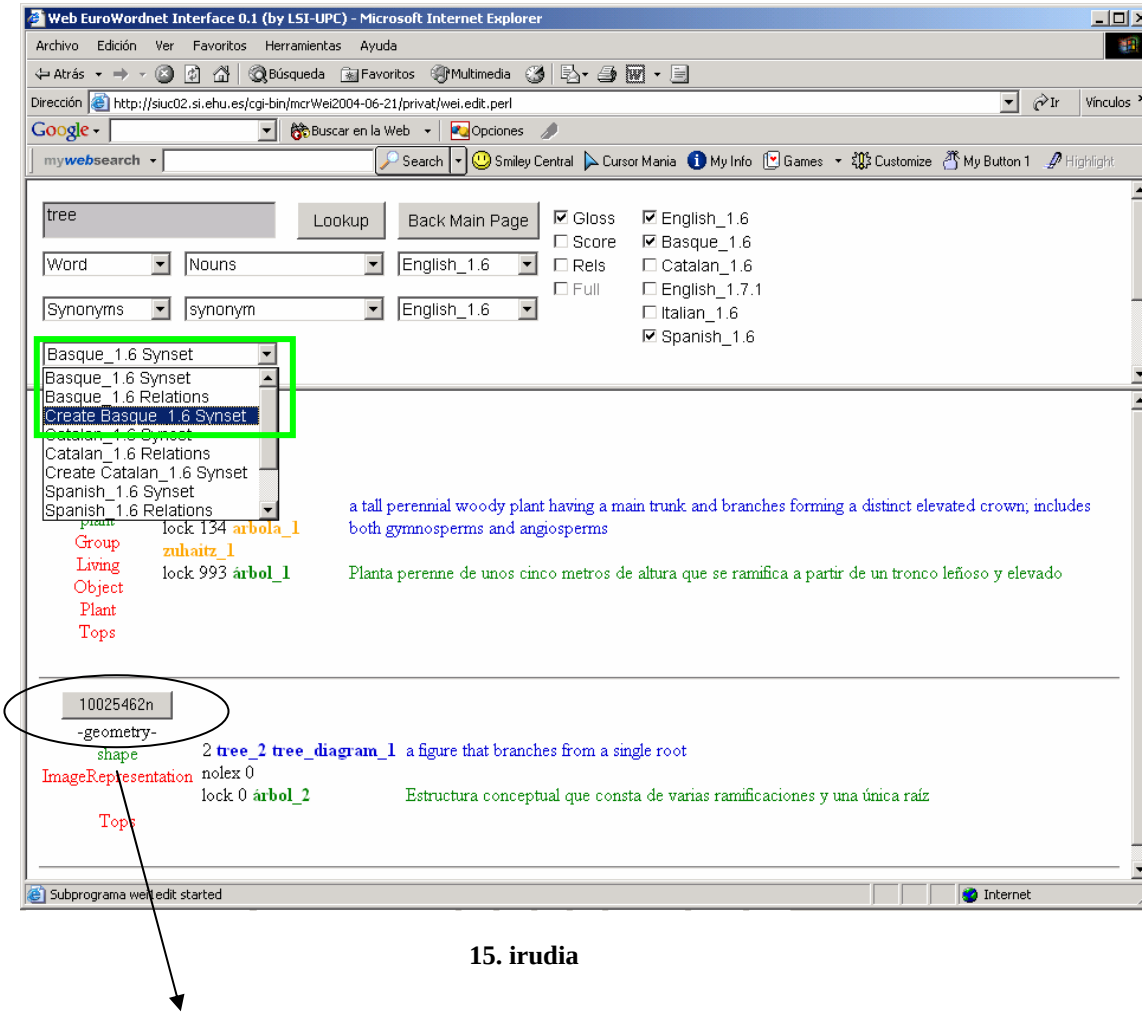
Bai *tree* hitzak, bai *árbol*-ek bi adiera dituzte EuroWordNet-en:

1. ‘landarea’
2. ‘diagrama’, ‘egitura’, ‘eskema’

Orduan, *zuhaitz* hitzak ere bigarren adiera hori (‘diagrama’) non txertatu baduela jakinda, synset horretan euskal ordaina sartuko da.

Aurreko irudian ikus daiteke 10025462 synset-ak ‘egitura’ edo ‘diagrama’ adiera duela. Orduan, editoreak synset-aren barruan sartu behar du, eta horretarako, **synset-zenbakiaren** gainean klikatu. Horrekin batera, kontuan izan behar dugu, synset horretan lehendik euskarako ordainen bat zegoen ala ez; aurretik synset-ean euskarako ordainik **egongo ez balitz**, 15. irudian borobilean markaturik agertzen den kutxatilan *Create Basque_1.6 Synset* aukeratu behar da euskal ordaina sartzeko. Aldiz, aurretik synset-ean euskarako ordainen bat **egongo balitz,** *Basque_1.6 Synset* aukeratu behar da[79].

[79] Aurretik euskal ordainen bat badago, hau egokia izan daiteke, eta egin nahi dena sinonimo bat gehitzea baino ez da. Horretarako, atal honetan aipatutako pausoak jarraituko ditugu. Dagoen ordaina okerra balitz, ezabatu beharko genuke, eta hori A.3.2.2.3 atalalean dator azalduta.

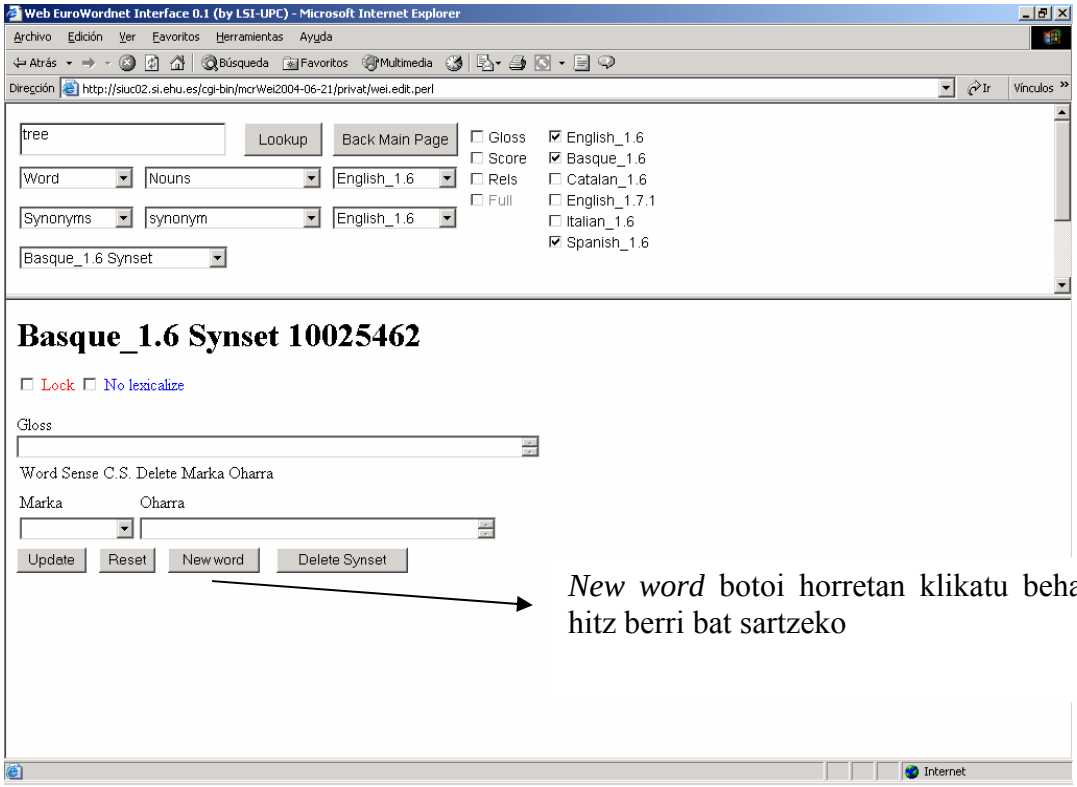


15. irudia

Synset-zenbakiaren gainean klikatu
eta leiho berri hau agertuko da

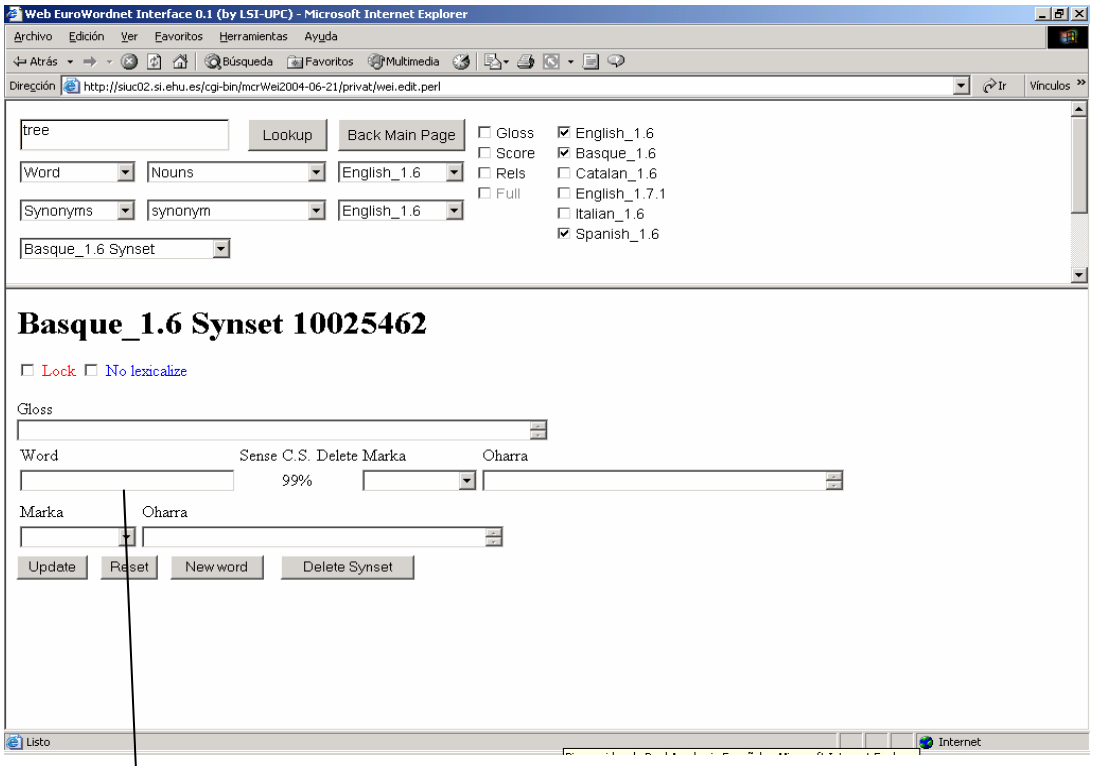
(*Leiho berri horretan agertuko diren leihotxo eta botoiak pixkanaka azalduz joango gara, behar ditugun neurrian alegia)

Ikus daitekeen bezala 10025462 synset-aren barruan gaude:



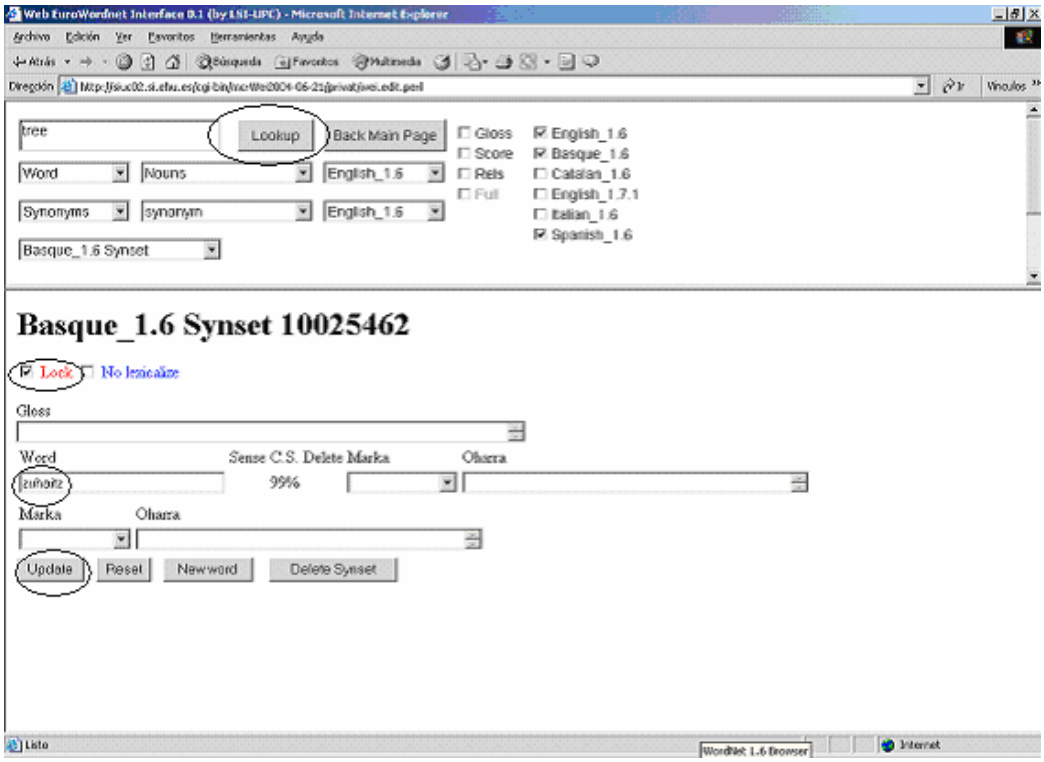
16. irudia

Ondoren, leihoak beste itxura bat hartuko du, eta ***zuhaitz*** hitza sartu ahal izango da:



17. irudia

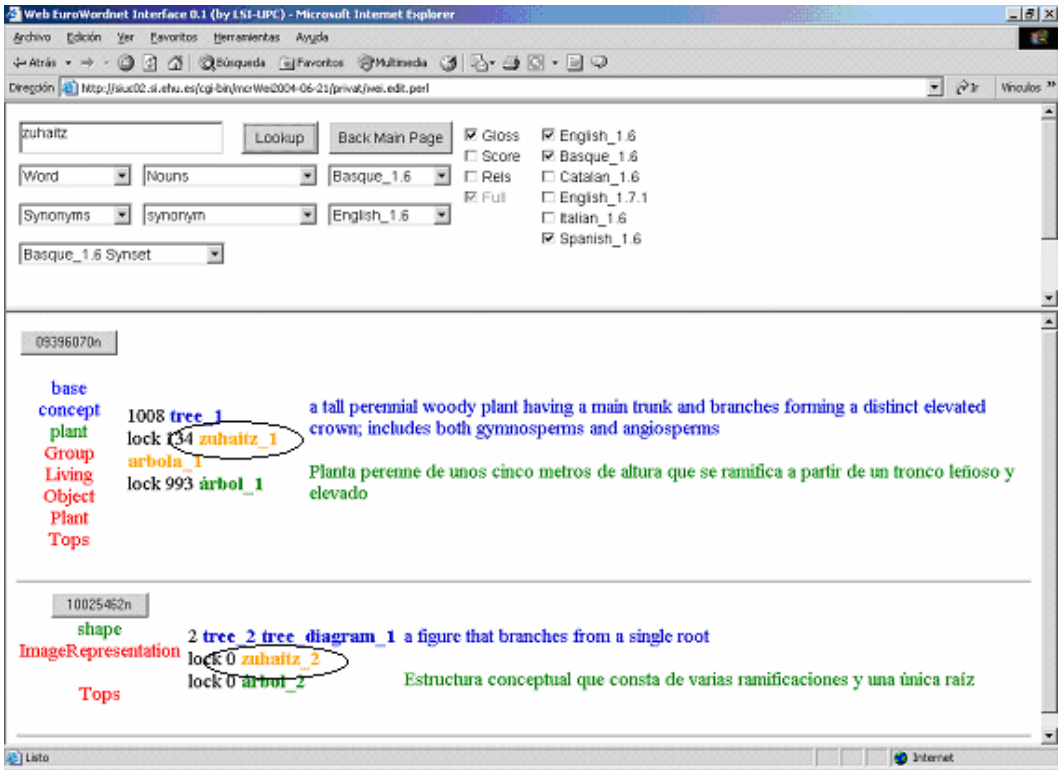
[80] HAULa bada, eta marratxorik ez badu (*polizi agente)*, bi osagaien artean "_" ikurra gehitu behar zaie: *polizi_agente*. Bestela, sistemak ez du hitz hori ezagutuko.



18. irudia

Irudian ikusten den bezala *zuhaitz* sartuta dago, eta eragiketa bukatzeko ***Lock*** marka jarri (eskuz landuta dagoela adierazteko) eta ***Update*** botoia sakatu behar da (synset-an egindako azkeneko aldaketak eguneratzeko).

Beraz, orain EuskalWordNet-en *zuhaitz* hitzaren bilaketa eginez gero (***Lookup*** botoia sakatuz gero), bi synset dituela ikusiko da:

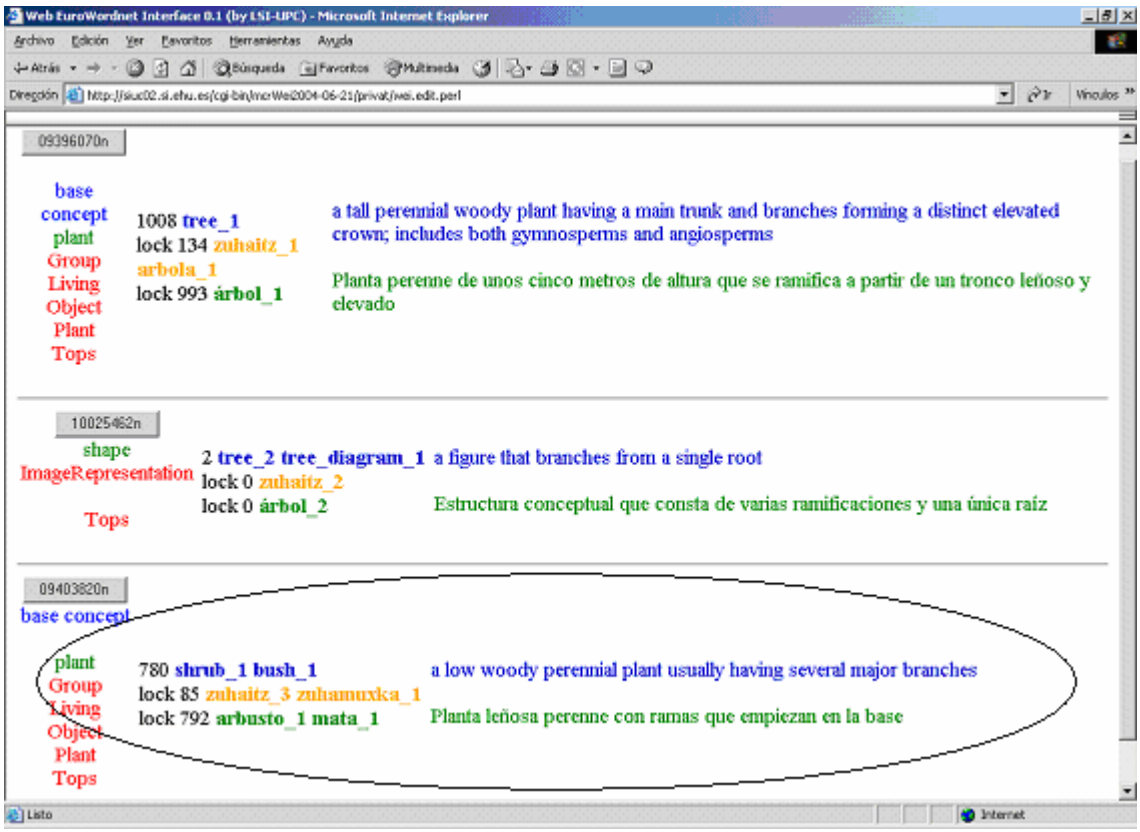


19. irudia

A.3.2.2.3 *Nola ezabatu euskarako ordaina synset batean*

Alderantzizko kasua gerta liteke: editorea hitz bat EuskalWordNet-en orrazten ari denean, ikus dezake ageri diren synset-en artean baten bat egokia ez izatea. Honek esan nahi du adiera hori ez dagokiola lantzen ari den hitzari. Bestela esanda, hitzaren adiera zuzenen artean ez dagoela synset horrek adierazten duena. Beraz, editoreak hitza ezabatu behar du synset horretatik. Adibide bat ikusiko dugu:

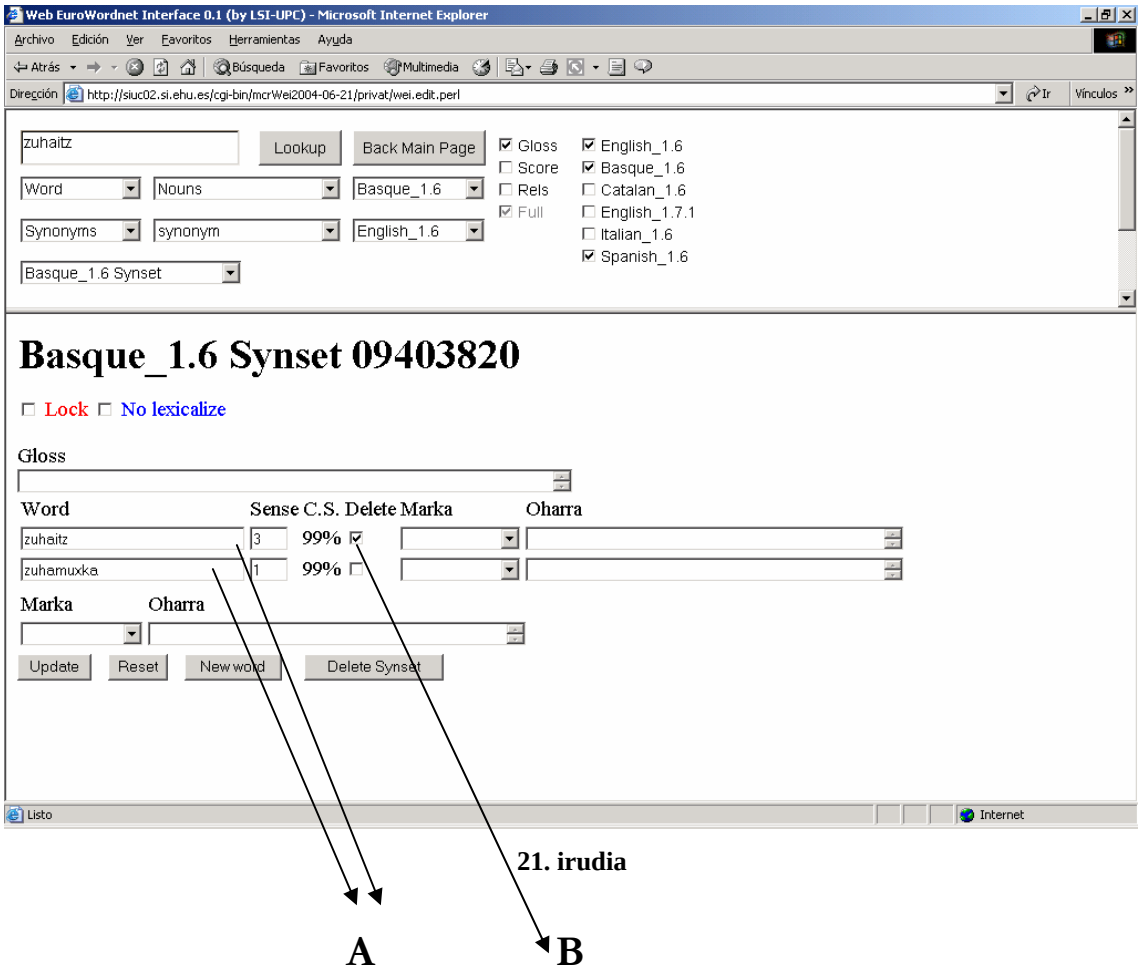
Demagun, EuskalWordNet-en *zuhaitz* hitzaren kontsulta egitean, ondorengo emaitza agertzen dela:



20. irudia

Irudi honetan *zuhaitz* hitzak hiru synset dituela ikusten da. Lehenengo biak aurreko ataletan landuak izan dira, baina hirugarrena berria da.

Aurreko ataletan (A.3.2.2.2 puntuan zehar) ikusiriko pauso guztiak jarraitu ondoren ondorio honetara iritsiko gara: *zuhaitz* hitzak ez du bere adieren artean gaztelaniaz *arbusto* edo *mata* dutenak, eta ingelesez *shrub* edo *bush* hitzek dutena. Adiera horretarako egokia da synset berean dagoen beste variant-a: *zuhamuxka*. Beraz, editoreak *zuhaitz_3* variant-a synset-a horretatik ezabatuko du. Horretarako, euskal ordain bat sartzeko bezala (A.3.3.2.2 atalen azalduta dagoen bezala) synset-aren barruan egin behar dira aldaketak. Orduan, editoreak synset-zenbakiaren gainean klikatu behar du, adibide honetan 09403820 synset-zenbakian. Gainera, kasu honetan, synset-ak badu euskarako ordainen bat, beraz, A.3.3.2.2 atalean esan bezala, *Basque_1.6 Synset* aukeratu beharko da (ikus 15. irudia). Ondoren, berriro, leiho hau agertuko da:



21. irudia

A = synset horrek dituen bi variant-ak

B = editoreak ezabatu nahi duen variant-ean ***Delete***-ren azpian dagoen laukitxoa markatu behar du.

Ondoren, aurrekoan bezala, ***Lock*** laukitxoa markatu behar du (eskuz landuta dagoela adierazteko) eta ondoren ***Update*** (egin den aldaketa eguneratzeko). Azkenik, ***Lookup*** botoia sakatzen bada, EuskalWordNet-ek *zuhaitz* hitzaren bilaketa egingo du eta ikusiko da bi synset-ekin geratu dela.

A.3.2.2.4 *Variant guztien orrazketa*

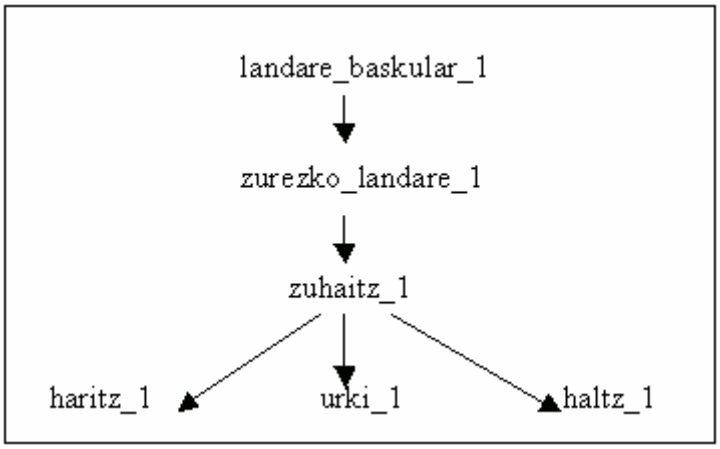
Orrazketaren beste zereginetako bat synset-eko beste variant-ak aztertzea izango litzateke. Bide batez, gainontzeko euskal variant-ak (baleude) zuzenak diren ere aztertu behar du editoreak: *zuhaitz_1*-en kasuan, ageri da beste variant bat: *arbola_1*. Eta hiztegiek erakusten digute *arbola* hitza *zuhaitz*en sinonimoa dela, eta berdin erabil daitezkeela. Beraz, synset-ean utziko litzateke. Bigarren synset-aren kasuan, ez da beste variant-ik agertzen, beraz, ez dago variant-ik aztertzeko.

Orduan, zeregin honetarako editoreak euskal hiztegietara jo beharko du (elebakar eta elebidunetara), synset horretan agertzen diren variant guztiak sinonimoak

diren egiaztatzeko. Baten bat egokia ez balitz, ezabatu beharko luke (ikus A.3.2.2.3 atala). Eta alderantziz, beste aukeraren bat aurkituko balu, gehitu beharko luke (ikus A.3.2.2.2 atala).

A.3.2.2.5 *Hiperonimo eta hiponimoen orrazketa*

Azkenik, synset bakoitzaren euskal hiperonimo eta hiponimoen hierarkia egokia den birpasatu beharko du editoreak.



22. irudia

Beste zenbait gauzen artean, editoreak arretaz aztertu behar du hitz batean hiperonimoan eta hiponimoan hitz bera (variant bera) ez agertzea. Gure adibidean ez da gertatzen, baina oso arrunta izango da beste hitz batzuen kasuan. Horrelako kasuak hurrengo atalean (A.3.4.1.3) landuko dira, eta bertan nola jokatu jakiteko irizpideak aurkitu ahal izango dira

A.3.2.3 *Orrazketaren zalantzak eta arazoak: irizpideak*

Aurreko atalean azaldu dugun prozesuan, hau da, hitz baten lanketan, askotan sortu dira hainbat arazo eta kasu berezi: euskaraz lexikalizatu gabeko synset-ak, kategoria bateraezinak, bereziki landu beharreko hitzak, adiera orokorregiak edo espezifikoegiak, eta beste zenbait zalantza eta arazo. Orain arte, zalantza hauek guztiak editoreak zerrenda batzuetan sailkatzen zituen, baina zalantza sortzen duten hitz hauek EuskalWordNet-en landu ahal izateko, zerrendetako zalantzak bildu eta aztertu dira, erabaki batzuk hartuz, eta irizpide batzuk finkatuz. Horrekin batera, editorearentzako beharrezkoak ziren zenbait aldaketa egin dira interfazean, eta horiek azalduko ditugu ondoko irizpide eta adibideetan.

A.3.2.3.1 *Nolex kasuak*

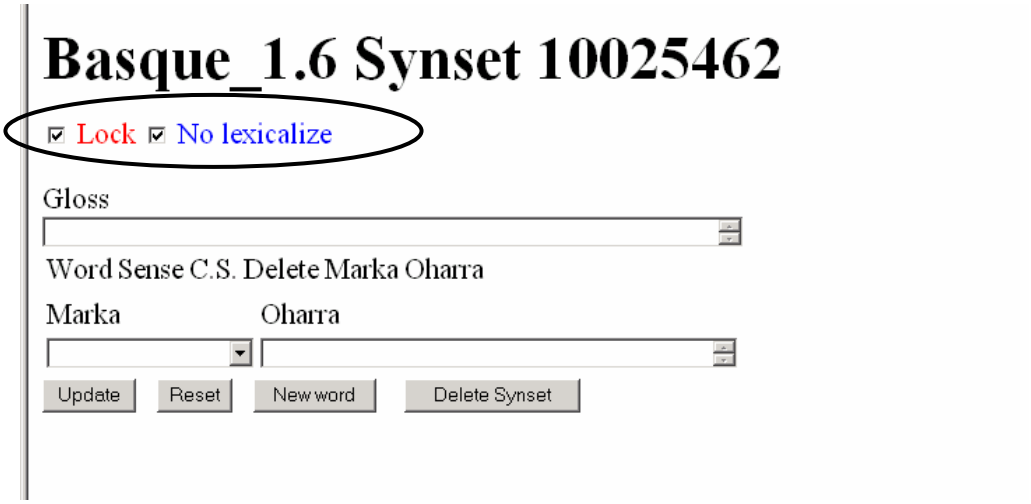
Atal honetan, euskaraz lexikalizaturik gabeko kasuak aztertuko ditugu. Noiz gertatzen da? Beste hizkuntza batean lexikalizaturik dagoen synset batek euskaraz ordainik ez duenean; hau da, gure hizkuntzan synset hori adierazteko esamolde edo esapide batera jo behar dugunean. Orduan, synset hori ***No lexicalize*** dela esaten da (aurrerantzean *Nolex)*, eta ikusiko dugun bezala, marka hori jartzen zaio[81]. Ondoren, *Nolex* kasu desberdinak aztertuko ditugu.

A.3.2.3.1.1 *Nolex arrunta*

Nolex arrunta ingeleseko[82] synset-ak euskaraz ordainik ez duenean gertatzen da, hau da, synset horren adiera euskaraz lexikalizatuta ez dagoenean. Adibidez:

forties: *the time of life between 40 and 50*.

Euskaraz ez dago hitzik synset hori adieraz dezakeenik, hots, euskaraz kontzeptu hori ez dago lexikalizatuta. Beraz, editoreak horrelako kasuetan synset horren barruko interfazean ***Nolex*** eta ***Lock*** marka jarriko dizkio, eta synset hori euskal variant-ik gabe uzten da.



23. irudia

[81] *Nolex* marka daramaten synset-en EuskalWordNet-en interfaze publikoan ez daude ikusgarri, hau da, interfaze pribatutik bakarrik hel daiteke hauetara.

[82] Esan bezala, EuskalWordNet EuroWordNet-en oinarrituta dago. Bertan hainbat hizkuntza daude (ingelesa, gaztelania, katalana, italiera...) baina oinarri gisa ingelesa hartuko dugu, nahiz eta beste hizkuntzak (batez ere, gaztelania) lagungarriak izan daitezkeen arren.

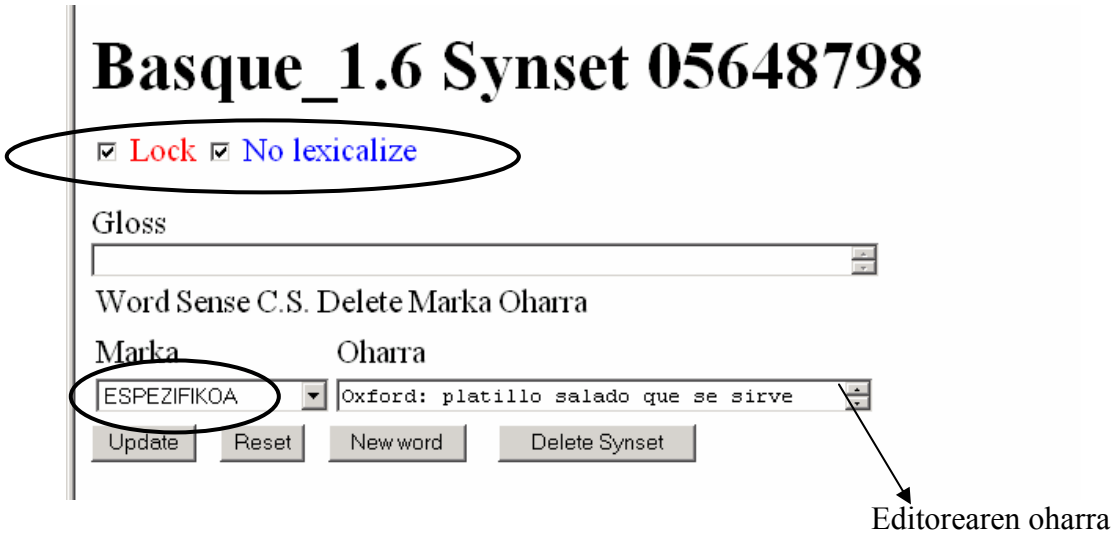
A.3.2.3.1.2 Espezifikoa Nolex

Badira beste hizkuntzetako zenbait synset oso adiera espezifikoa dutenak, eta nahiz eta, behar bada, euskaraz ordainen bat izan, ordain hori topatzea zaila gerta daiteke, eskura ez ditugun hiztegi espezializatuetara jotzea behartzen gaituelako. Adibidez:

False mistletoe: *American plants closely resembling Old World mistletoe*

Savoury: *(British) an aromatic or spicy dish served at the end of dinner or as an hors d´oeuvre*

Horrelakoetan editoreak ahal duen neurrian euskarako ordaina bilatzen saiatu behar du, orain arte aipatutako hiztegietan (ikus A.3.1 atala). Aurkituko balu, dagokion synset-ean sartuko luke. Baina ordainik topatuko ez balu, ***Espezifikoa*** eta ***Nolex*** markak jarriko ditu. Gainera, synset-aren ***Oharra*** laukitxoan eman dituen pausoak idatzi behar ditu. Azkenean ***Lock*** markatuko du landu duela adierazteko[83].



24. irudia

A.3.2.3.1.3 Orokorra Nolex

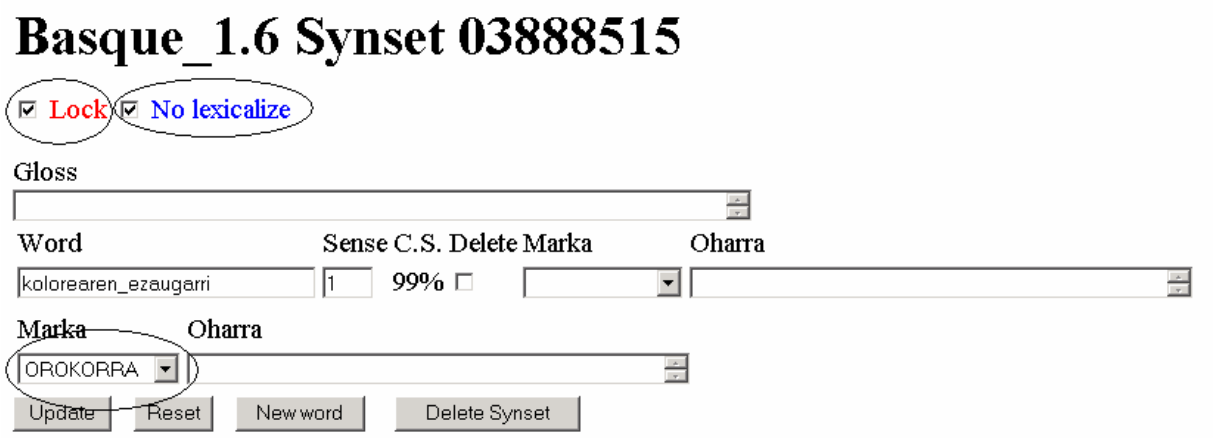
Ingeleseko WordNet-en kontzeptu orokor batzuk izendatzeko terminoak "asmatu"-edo egin dira. Adibidez, *entity* azpian daudenean *imaginary place, body of water, unpleasant woman,* eta halakoak, hauen baitan dauden hiponimoen sailkapena errazteko sortu dira. Beste hitz batzuetan esanda, synset hauek "antolatzaileak" direla esan daiteke, hiponimo sorta bat izendatzeko beharrezkoak. Horregatik, nahiz eta kontzeptu hori berez lexikalizatua ez egon, adierazi egiten da hierarkia ulergarriagoa egitearren.

[83] Izen berezi batzuk (bataila batzuen izenak, besteak beste) era honetara marka daitezke.

Entity, physical thing

- *imaginary place*
- *body of water, water*
- *enclousure, natural enclousure*

Horrelakoak euskaratzean, editoreak saiatu behar du ahal duen neurrian euskarako ordaina topatzen. Aurkitzen badu, synset-ari lotuko dio. Baina aurkitzen ez badu, ***Orokorra*** eta ***Nolex*** bezala markatu ditu; eta ***Oharra*** eremuan hartutako erabakiaren berria emango da (zer hiztegietan begiratu dugun eta abar). Bukatzeko ***Lock*** marka ere jarriko du.



25. irudia

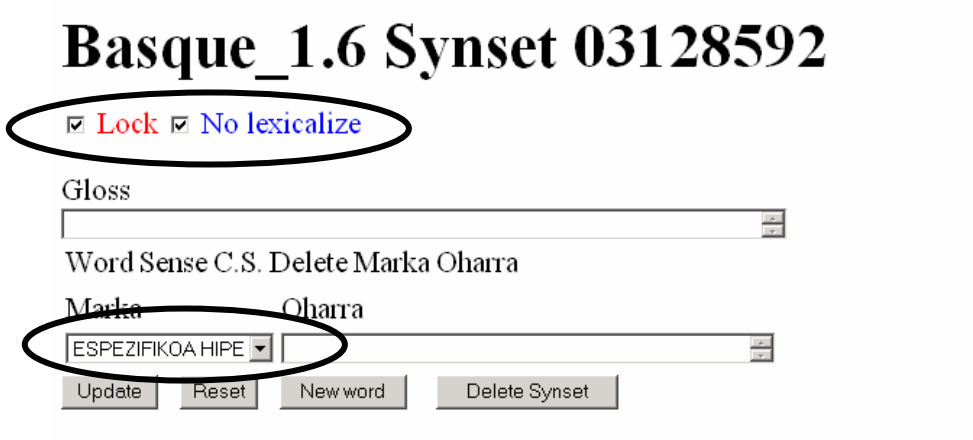
A.3.2.3.1.4 Hiperonimia eta Nolex

Esan bezala (ikus A.2.2.1 atala), hiponimoak hiperonimoen zehaztapenak dira. Dirudienez, ingelesak hiperonimia-hiponimia erlazio hau hitz desberdinekin adierazteko gaitasuna handia du, euskarak baino gehiago behintzat. Horregatik, askotan, ingeleseko hiponimoak euskaratzean hiperonimo eta hiponimoak hitz berarekin adieraz daitekeela gerta daiteke; bestela esanda, ingeleseko hiperonimo bate hiponimoetan bakoitzarentzat termino desberdin bat dagoenean, euskaraz hiperonimo eta hiponimo horiek hitz bera izango dute. Adibideetako bat *buru* hitza da:

- lock 10 **buru_9 mutur_14 punta_7**
  end_1 *either extremity of something that has length*
  - lock 5 **buru_27 mutur_7 punta_1**
    point_15 *sharp end*
  - lock 0 **buru_28 mutur_24 punta_13**
    pinpoint_3 *the sharp point of a pin*
  - *lock 0* ***buru_26 mutur_18 punta_8***
    tip_1 *the extreme end of something; especially something pointed*

Ikus daitekeen bezala, *buru_9* hiperonimoaren azpiko hiru hiponimoen variant-ak berdinak dira, hau da, denek *buru, punta* eta *mutur* gisa adierazita daude, euskaraz ez dira bereizten. Horrelakoetan editoreak jarraitu beharreko irizpidea honako hau da: hiponimoei ***Nolex*** marka jarri eta hiperonimoa bere horretan utzi. Hala ere, hau guztia

adierazteko hiponimoei marka bat jarriko zaie: ***Espezifikoa HIPE***, ***Nolex*** eta ***Lock*** bezala markatuko dira.



26. irudia

Oharra: Kasu honetaz ohartu ahal izateko, EuskalWordNet-en synset edo hitz baten kontsulta egitean, honen hiperonimoak eta hiponimoak beti eskatzea gomendagarria da. Bestalde, irizpide hau hiperonimo-hiponimo hurbilen artean bakarrik erabiliko da. Hala ez balitz, lan-taldeari ohartarazi behar zaio.

A.3.2.3.1.5 -TU/-TZE Nolex

Dirudienez, ingelesez *act* eta *state* eremu semantikoak oso ondo bereizten dituzte testuinguruaren arabera.

employment_1 the state of being employed or having a job ➔ **being employed**
employment_2 the occupation for which you are paid; "he is looking for a job"
employment_3 the act of giving someone a job ➔ **to employ**

Euskaraz, berriz, horrelako synset-ak topatzen ditugunean *-tu* eta *–t(z)e* atzizkiak behar ditugu. Adibidez, *employment_1* adierazteko, euskaraz *enplega**tu*** erabiliko litzateke:

10063469n

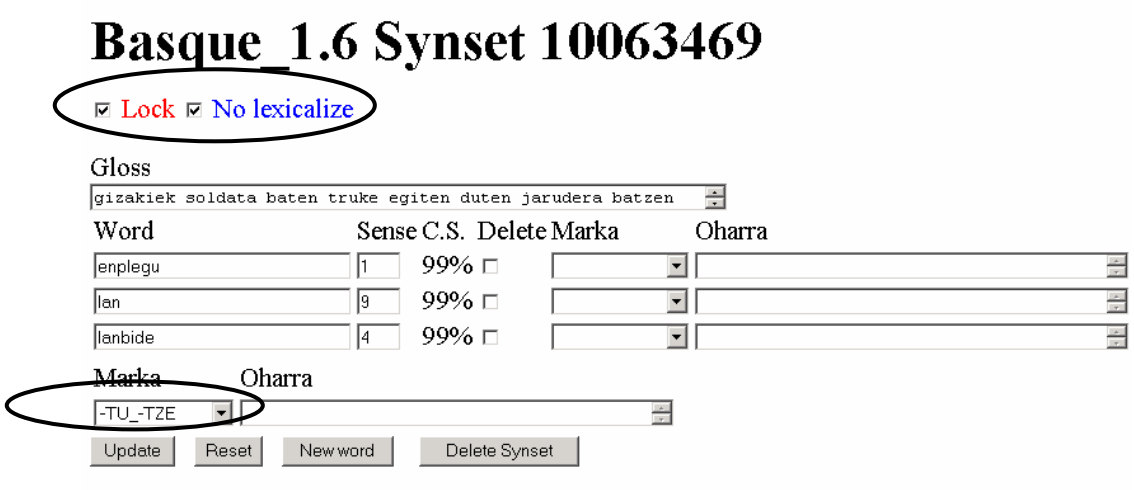
-	**10063469n** 0 **employment_1 employ_1**	the state of being employed or having a job
enterprise-	**10063469n** 0 **empleo_2 trabajo_11**	Hecho de tener un oficio
state	**ocupación_5**	gizakiek soldata baten truke egiten duten jarduera batzen
employs=	**10063469n** 0 **enplegatu_1**	duen sistema
Static=		

27. irudia

Baina, EuskalWordNet-en hiztegi-sarrerak txertatzen dira bakarrik, hau da, enplegatu hitzak, synset horrek duen adiera horrekin[84], hiztegietan ez du hiztegi-sarrerarik, eta oraingoz, horrelako synset-ei marka bat ezarriko zaie, hurrengo orrazketa

[84] *Enplegatu* hitzak hiztegietan bi hiztegi-sarrera izan ohi du; bata, aditza (*Ontsa enplegatu denbora*) eta bestea, izena (*Berrogei urtean enplegatu egon zen*). Bigarren adiera honek (*enplegatu* = ingeleseko *employee*), nahiz eta izena izan, ez du zerikusirik *employment_1* horrekin.

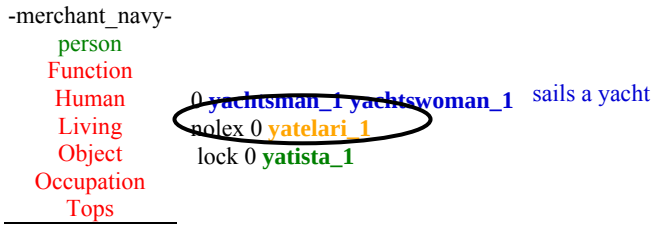
batean erreparatzeko. Beraz, kasu hauetarako ***–TU / -TZE*** marka sortu da. Hortaz, editoreak hitza (hiztegi-sarrera) sartuko du (*enplegu* eta ez *enplegatu*), jarraian, ***–TU / -TZE*** eta ***Nolex*** markatuko ditu, eta azkenik, synset-a ***Lock*** gisa utziko du:



28. irudia

A.3.2.3.1.6 Bestelako kasuak

Batzuetan, EuskalWordNet-en interfazea kontsultatzean, *Nolex* marka eta variant-a dituzten synset-ak topa ditzakegu. Adibidez:



29. irudia

Hauek orrazketaren beste fase batean egindakoak dira, gehienak, EuskalWordNet editatzeko irizpideak garatu gabe zeudenekoak dira. Egungo metodologia dela-eta, horrelako kasuak ez dira sortzen, baina horrelakoren bat topatuz gero, editoreak synset hori eskuliburu honetan zehaztutako irizpideen arabera moldatu beharko luke (nahiz eta synset-a ***Lock*** egon[85]). Hurrengo kasuistika gerta daiteke:

1. Variant-a hitz bat izatea (ikus 29. irudia): **hauek, normalean, *Nolex arrunta* eta *Espezifikoa Nolex* gisa tratatuko dira (ikus A.3.2.3.1.1 eta A.3.2.3.1.2 atalak).** Hala ere, kasuan kasu, irizpidea ezberdina izan daiteke –adibidez, hiperonimoan ordain bera agertzea (ikus ikus A.3.2.3.1.4 atala), edota synset horrentzat euskarako ordain apropos bat topatzea, eta abar.

[85] Nahiz eta synset hori landuta egon, baliteke aurreko orrazketako erabaki horrek txosten honetan azalduriko irizpideekin bat ez etortzea. Horregatik, errepasatzea komenigarria da.

2. Variant-a HAUL bat izatea:

-art-
-linguistics-
communication
Agentive
Communication
Manner
Mental
Purpose
Social
Tops
UnboundedEvent

2 **long-windedness_1 prolixity_1 prolixness_1 wordiness_1**
lock nolex 2 **hitzaldi_luze_1 kontakizun_luze_1 azalpen_luze_1** boring verboseness
lock 2 **prolijidad_1**

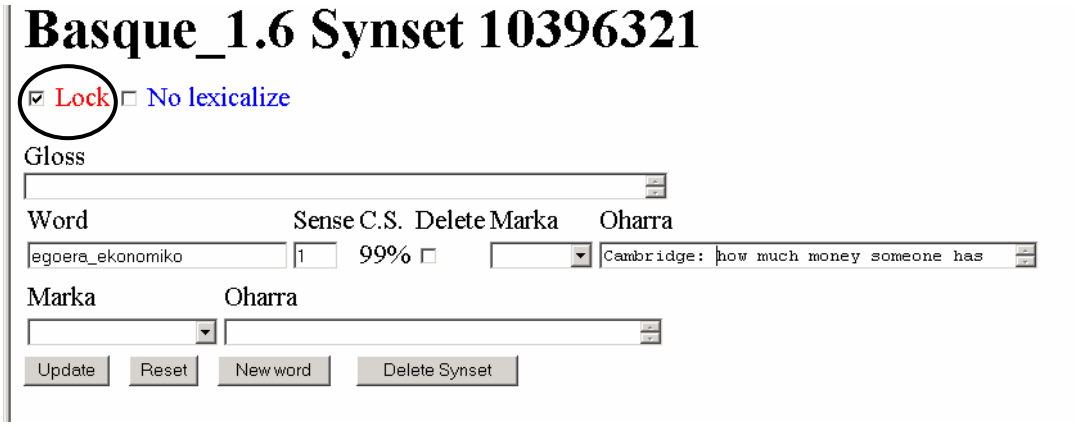
30. irudia

Esan izan dugun bezala, synset-en euskal ordainak eman ahal izateko, hiztegiak baliatzen ditugu, eta askotan, hiztegi elebidunak. Hala, ingeles/gaztelaniako hitz askok, euskaraz HAUL bat behar dute kontzeptu hori adierazteko. Adibidez, ingeleseko *prolixity* (gaztelaniako *prolijidad*) *Elhuyar Hiztegia*k *hitzaldi luze, kontakizun luze* eta *azalpen luze* bezala itzultzen ditu. Horrelakoak lantzeko orduan, ondorengo irizpideak jarraitu behar dira:

- *Euskal Hiztegia*n, *Euskal Hiztegi Modernoa*n, *Elhuyar Hiztegi* elebidunetan, *Euskaltermen* eta *EDBL*en HAUL hori **sarrera** edo **azpisarrera** bezala dagoen ala ez begiratu:

  a) Baldin badago (gutxitan gertatzen dena)**:** Adibidez*: egoera ekonomiko*

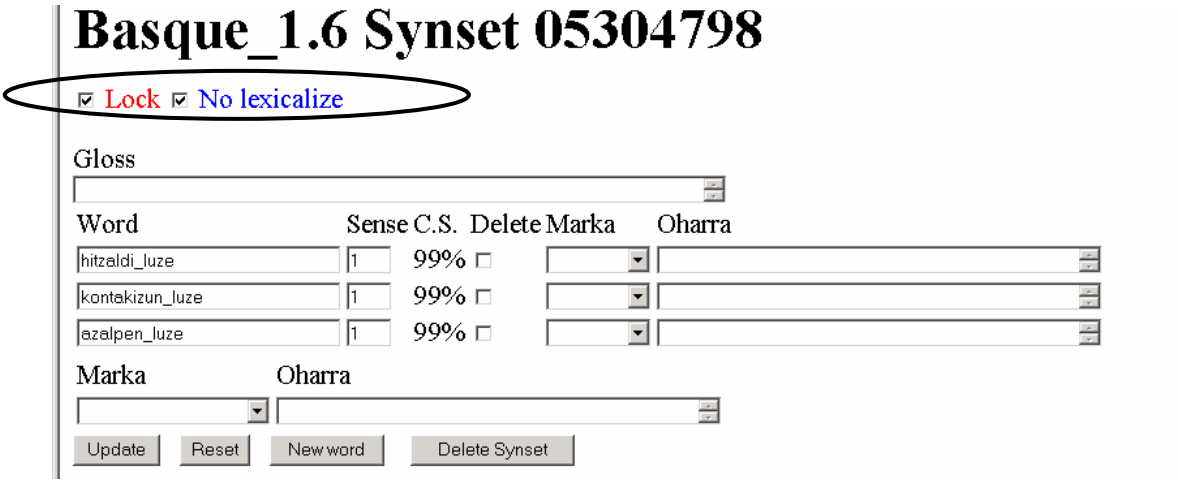
  - Variant-ari hitz anitzeko ordaina (*egoera_ekonomiko*) gehitu.
  - Synset-a ***Lock*** jarri.



31. irudia

b) Ez badago: Adibidez: *prolijidad: azalpen luze / cutis: aurpegiko larruazal*

  - Ordaina variant bezala jarri
  - Synset-a ***Nolex*** eta ***Lock*** jarri



32. irudia

Oharra: Synset hori adierazteko lexikalizatua (hiztegi-sarrera) den sinonimo bat baldin badago, orduan, HAULa ez da synset-ean gehituko. Adibidez, gaztelaniko *desagrado* hitzaren itzulpenak *Elhuyar*en *disgustu, desplazer; atseginik ez(a)* dira. Horien artean HAUL bat dago (*atseginik ez(a)*). Beste sinonimoak (*disgustu* eta *desplazer*) lexikalizatuak daudenez eta hiztegi-sarrerak direnez, EuskalWordNet-en txertatuko dira; *atseginik ez,* aldiz, hiztegi-sarrera bat ez denez, eta horren ordez lexikalizatuak dauden beste ordainak badaudenez, ez da variant gisa gehituko[86].

A.3.2.4 Variant-ei dagozkien kasuak

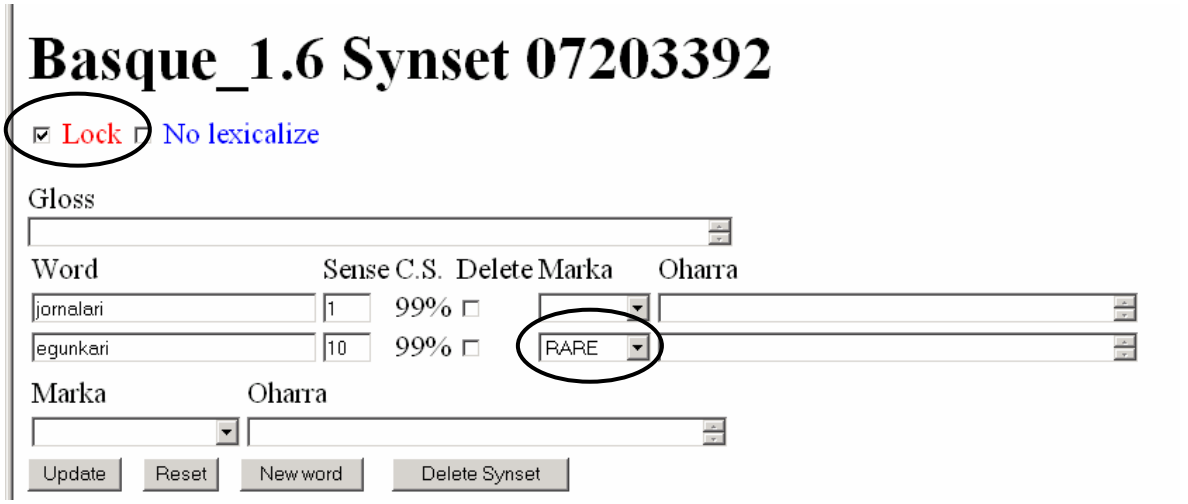
A.3.2.4.1 RARE marka

Euskalkietako aldaera desberdinekin arazoak sortzen dira zenbaitetan. Honako adibidea argia da *egunkari* izena. Hiztegietan gaztelaniako *periódico* adieraz gain, iparraldean badu beste adiera bat: *jornalero*. Hala, editoreak *jornalari* kontzeptua lantzean, baliteke synset-en batean *egunkari* hitza topatzea edota txertatzeko zalantza izatea. Horrelakoetan, editoreak jarraitu beharreko irizpideak hauexek dira:

- Hitz horiek **ez dira** EuskalWordNet-en sartuko:
  - EDBLn ***RARE*** markadunak direnean
  - *Elhuyar Hiztegi Handian* eta *Txikian*, *Euskal Hiztegi Modernoa*n, *Hiztegi Batua*n eta *Euskal Hiztegia*n gutxi erabiliak eta zaharkituak bezala agertzen direnean.

[86] Irizpide hau *mota* ezberdinak adierazten duten synset-ekin ere erabil daiteke. Esate baterako, *irakasle* kontzeptuaren azpian *irakasle motak* (*musikako irakasle, matematikako irakasle...*) egongo dira hiponimo gisa; horrelakoak lantzea ala ez, editorearen esku dago, baina variant-a sartuz gero *Nolex* markatu beharko du.

- **<u>Dagoeneko horrelako hitzen bat EuskalWordNet-en badago</u>**, ***RARE*** marka jarriko zaio variant-ei, eta synset-a ***Lock*** geratuko da.



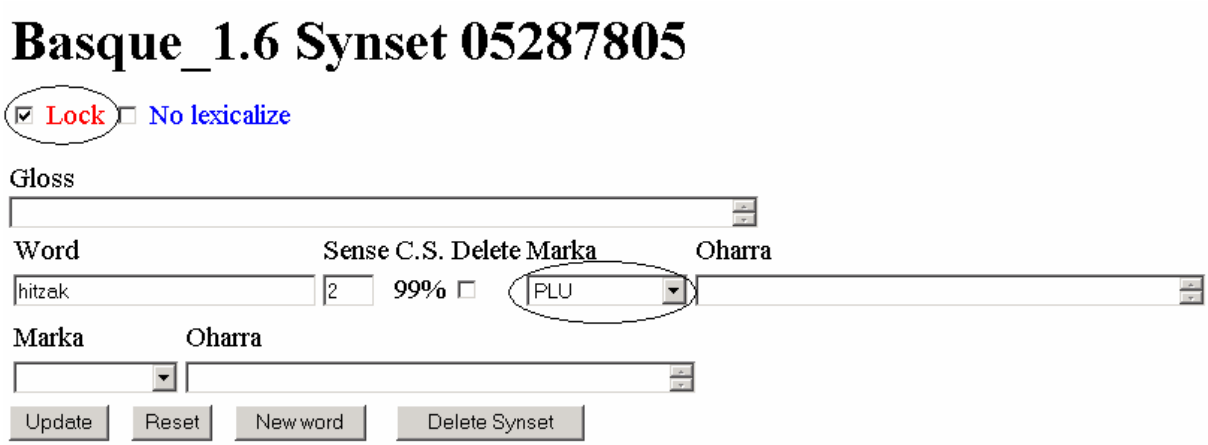
33. irudia

A.3.2.4.2 *<u>PLU marka</u>*

Zenbait synset-etan gerta liteke euskal ordainaren erabilera beti plurala izatea[87]. Adibidez:

hitzak = 'abestien letra'
paperak = 'dokumentazioa'

Kasu hauetan editoreak pluraleko forma horiek (*hitzak, paperak*) synset-ean lotuko ditu eta ***PLU*** marka jarriko die. Ondoren, ***Lock*** geratuko da synset-a.



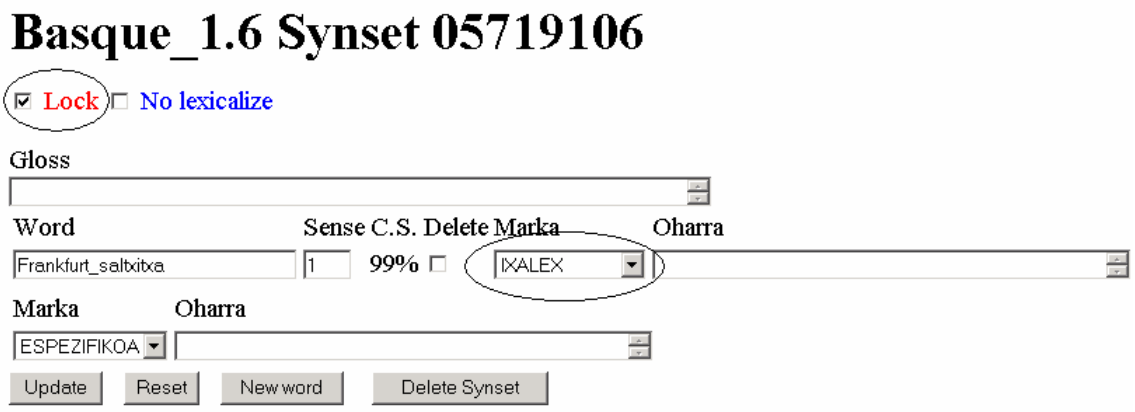
34. irudia

[87] Kasu hau, etiketatze-lanean ere (5.2.1.3 atalean) aipatu dugu.

A.3.2.4.3 *IXALEX marka*

Synset baten euskal ordainaren bilaketan, gerta daiteke hiztegietan ez topatzea. Hala ere, editorea ziur badago adiera horrentzat badela hitz bat lexikalizaturik dagoena, sartu egingo du synset-ean eta ***IXALEX*** marka jarriko dio, ondoren ***Lock*** utziko du. Adibidez: *Frankfurt saltxitxa, egun siderial, irudi bidimentsional...*

Askotan, editorearen intuizioa hiztegiengatik eragina izan dezake. Hau da, *two-dimensional_figure* lantzen ari bada, eta hiztegietan *two-dimensional, bidimentsional* gisa itzultzen badute eta hiztegiko adibidea *esfera bidimentsional* bada, pentsatzekoa da, adjektibo hau *irudi*rekin batera ere erabil daitekeela.



35. irudia

Kasu honetan, gainera, synset-ak espezifiko marka dauka.

A.3.2.5 *Idazkera zalantzak*

A.3.2.5.1 *Marratxodun hitzak*

Herri-, *haur-* eta bezalako izenek marratxoa daramatenean, hau mantendu egingo da, hau da, hitz batek berarekin beti marratxoa eskatzen badu, orduan, hitzarekin batera marratxoa txertatuko da EuskalWordNet-en[88].

A.3.2.5.2 *Artikulua daramaten hitzak*

Kasu hauetan, editoreak jarraitu behar duen irizpidea ***–a*** kentzea da. Beraz, hiztegietan *atseginik ez(a)* bezalakoak aurkitu arren, EuskalWordNet-en *atseginik ez* txertatuko da.

Hala ere, horrelako HAULekin kontuz ibili beharra dago, ikus A.3.2.6.6 atala.

[88] Atal hau adjektiboekin harremanetan dago. Oraindik adjektiboak txertatzen hasi ez arren, izenak lantzean horrelako arazoak aurreikusi egin dira. Hala ere, honi buruz A.3.2.7.1 atalean mintzatuko gara.

A.3.2.5.3 *Idazteko era desberdinak*

Gerta liteke, hitz berak aukera bat baino gehiago izatea idazteko orduan, eta hauek hiztegietan jasota egotea. Adibidez:

polizi agente
polizia-agente
agente

Editoreak EuskalWordNet-en idazteko era guztiak sartuko ditu[89].

A.3.2.5.4 *Hizki larriak eta xeheak*

Gerta daiteke, hitz bera batzuetan hitz larriz eta besteetan letra xehez agertzea hiztegi eta dokumentu desberdinetan. Orduan, editoreak EDBL datu-base lexikalera joko du; eta bertan agertzen dena izango da irizpide erabakia hartzeko. Adibidez: *Jainko* ala *jainko*. Kasu honetan EDBLk biak jasotzen ditu biei buruzko informazio zehatza ematen du.

A.3.2.6 *Bestelako zalantzak*

A.3.2.6.1 -keta, -kuntza, -mendu*... bezalako sinonimoak*

Mota horretako atzizkiak dituzten hitzen artean sinonimia gertatzen da sarritan. Adibidez:

antolaketa
antolakuntza
antolamendu

Editorearentzako irizpidea honakoa da: *Elhuyar Hiztegi Txiki*ko hiztegi-sarrera gisa agertzen diren neurrian sartuko dira, hau da, synset batean *antolaketa* gehitu nahi badugu, eta *Elhuyar Hiztegi Txikia*n hiztegi-sarrera gisa *antolakuntza* ere badago, orduan biak gehituko dira synset horretan. *Elhuyar Hiztegi Txiki*an *antolakuntza* egongo ez balitz, ez genuke gehituko.

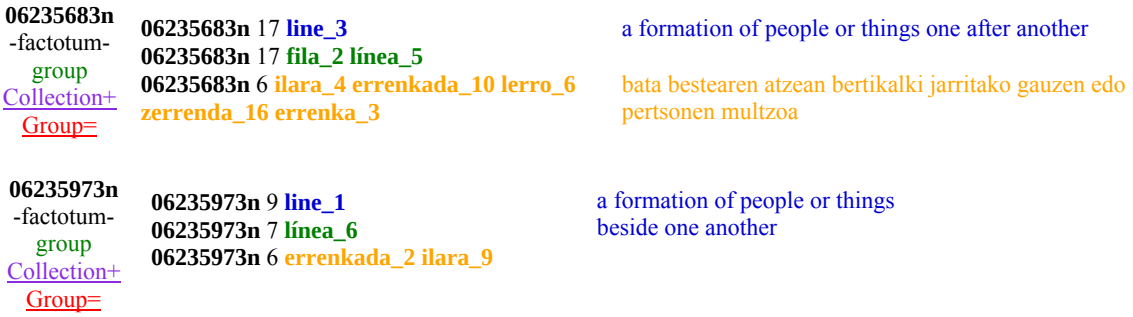
A.3.2.6.2 *Hiztegiak bat ez datozenean*

Batzuetan editoreak hiztegi desberdinetara jotzean, bateragarria ez den informazioarekin topa daiteke. Adibidez, gaztelaniako *salsera* txertatu nahi dugu EuskalWordNet-en. *Euskal Hiztegi Modernoa*n eta *Elhuyar Hiztegia*n begiratuz gero, itzulpen gisa *saltsaontzi* ematen du eta, *Euskaltermek* aldiz, *saltsontzi*. Euskaltzaindiak ez badu horri buruzko araurik, orduan, editoreak *uskal Hiztegi Modernoa*k eta *Elhuyar Hiztegia*k dioena jarraituko du.

[89] Atal honek HAULekin (geroago datorren A.3.2.6.6 atalarekin, alegia) harremanetan dago. Beraz, HAULak lantzean idazkera kontua izan beharrekoa da.

A.3.2.6.3 *Antzeko synsetak bereizteko zailtasuna*

Batzuetan oso antzekoak diren synset-en artean bereiztea oso zaila gertatzen da. Adibidez, *ilara* hitzaren kasuan, hurrengo bi synset-ak ditu, eta euskaraz horiek nekez bereiz daitezke:



36. irudia

Kasu honetan bi synset-ak ingeleseko *formation* synset-etik datoz:

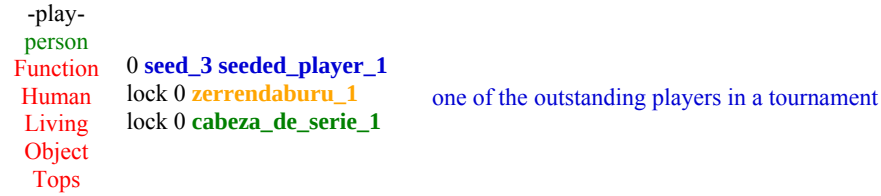
=>formation (an arrangement of people or things acting as a unit)
=> line (a formation of people or things one after another; "the line stretched clear around the corner")
=> line (a formation of people or things beside one another; "the line of soldiers advanced with their bayonets fixed"; "they were arrayed in line of battle")

Ingeleseko *formation* euskaraz *ilara* itzuli ahal izango balitz, A.3.2.3.1.4 ataleko kasuaren (*buru*rena, alegia) berdina litzateke; baina, oraingoan, ezin dira bi synset hauek ***Nolex*** utzi hiperonimoari *ilara* jarriz (*formation* ez baita euskarako *ilara*). Hortaz, horrelako synset-ak lantzean, maila bereko synset-ak direnak, polisemikotzat joko dira, hots, *ilara* hitzak gutxienez EuskalWordNet-en bi synset horiek izango ditu[90].

Bestalde, horrelako arazoen aurrean, ingelesekoWordNet 2.0 bertsioa kontsultatzea komenigarria da, 1.6 bertsiotik 2.0 bertsiora zuzenketak/aldaketak egon daitezkeelako.

A.3.2.6.4 *Adieren egokitasuna*

Gerta daiteke ingeleseko synset-a eta euskarakoa erabat baliokideak ez izatea. Adibide argia hauxe dugu: *zerrendaburu*



37. irudia

[90] EuSemCor etiketatzean, horrelako synset-ak bateratu daitezke, hau da, agerpen bati bi etiketa (*ilara_4* eta *ilara_9*) ematea badago, hauek testuinguruan bereiz ezinak baitira.

Euskarako *zerrendaburu*k esanahia zabalagoa du, ez du bakarrik kiroleko adiera ingelesez bezala; esaterako, politikan hauteskundeetarako zerrendetan *zerrendaburu* hitza ere maiz erabiltzen da. Ingelesez, berriz, *seed* eta *seed player* kiroletarako erabiltzen dute soilik. Beraz, adierak ez dira erabat baliokideak.

Kasu hauetan editoreak honela jokatu beharko du: hauen guztien hiperonimoa *zerrendaburu* balitz, hiperonimoari gehituko litzaioke variant hau eta hiponimoak ***Nolex*** bezala utziko litzateke (ikus eranskinaren A.3.2.3.1.4 atala). Baina, hau ez da kasua, eta ingeleseko WordNet-en ez dago *zerrendaburu* orokor hori adierazten duen kontzepturik. Beraz, euskarako *zerrendaburu* polisemiko bezala landuko da, hau da, adiera bat baino gehiago dituen hitz baten gisa.

Bestalde, horrelako arazoen aurrean, ingeleseko WordNet 2.0 bertsioa kontsultatzea komenigarria da, 1.6 bertsiotik 2.0 bertsiora zuzenketak/aldaketak egon daitezkeelako.

A.3.2.6.5 Figuratiboak

Zenbait kasutan izen batek adiera figuratibo/metaforikoren bat izan dezake. Ingeleseko WordNet-en horrelako batzuk jasota daude:

-factotum-
cognition
Mental
Static

0 **teacher_2** a personified abstraction that teaches "books were his teachers"

38. irudia

Horrelakoak, batzuetan euskarara itzuli daitezke eta beste batzuetan, ordea, ez. Hau da, gerta daiteke, ingeleseko hitz horrek (*teacher*) duen adiera figuratiboa euskarako ordainak (*irakasle*) ere horixe bera izatea. Horrela bada, editoreak synset horretan euskarako ordaina txertatuko du[91]:

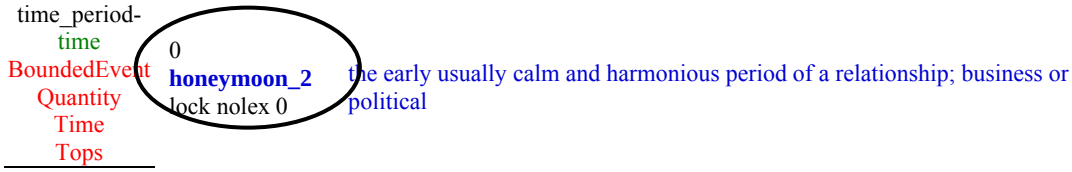
-factotum-
cognition
Mental
Static

0 **teacher_2**
lock 0 **irakasle_5** a personified abstraction that teaches "books were his teachers"
nolex 0 irakasten duen abstrakzio pertsonifikatua “nire irakasleak liburuak izan ziren”

39. irudia

Aldiz, euskarak ordain hori figuratibo gisa izango ez balu, editoreak synset hori ***Nolex*** eta ***Lock*** utziko luke. Esate baterako, ingeleseko *honeymoon* izenak beheko synset-eko adiera figuratiboa du. Baina, euskaraz itzulpen zuzena den *eztei-bidai* izenak ez du adiera hori. Horregatik, beheko synset-ean ez dago euskarako variant-ik, eta synset-a ***Nolex*** eta ***Lock*** gisa utzi da.

[91] Erabaki hori hartzeko, euskaraz izen bat figuratibo gisa erabiltzen dela egiaztatzeko, editoreak hiztegi eta corpusetara jo beharko du.



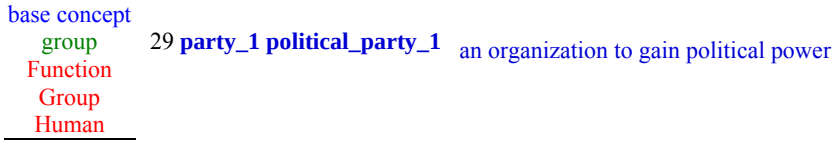
40. irudia

A.3.2.6.6 HAULak

5.2.1.3 atalean HAULak aipatu ditugu, eta bertan esan bezala, batzuetan HAUL batek adierazten duen adiera bera, HAULeko osagai **bakar batek ere** adieraz dezake (elipsiaren antzeko zerbait gertatzea, alegia):

- ***Partidu politiko*** *guztiek uka dezatela...*
- ***Partidu*** *guztiek uka dezatela...*

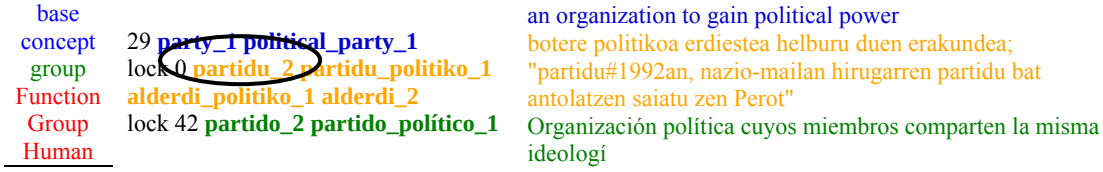
Ingeleseko WordNet-en HAULak ere badaude:



41. irudia

Editoreak synset honetan *partidu politiko* txertatuko luke. Baina gainera, *partidu* hitzak HAUL hori adieraz dezakeen ala ez egiaztatu beharko luke. Horretarako, hiztegi eta corpusetara jo beharko du. Egoera horren aurrean bi aukera egon daitezke:

a. Hiztegi edota corpusetan hori egiaztatzen bada, editoreak *partidu* hitza ere synset horretan sartuko du (A.3.2.3.1 ataleko pausoak jarraituz):



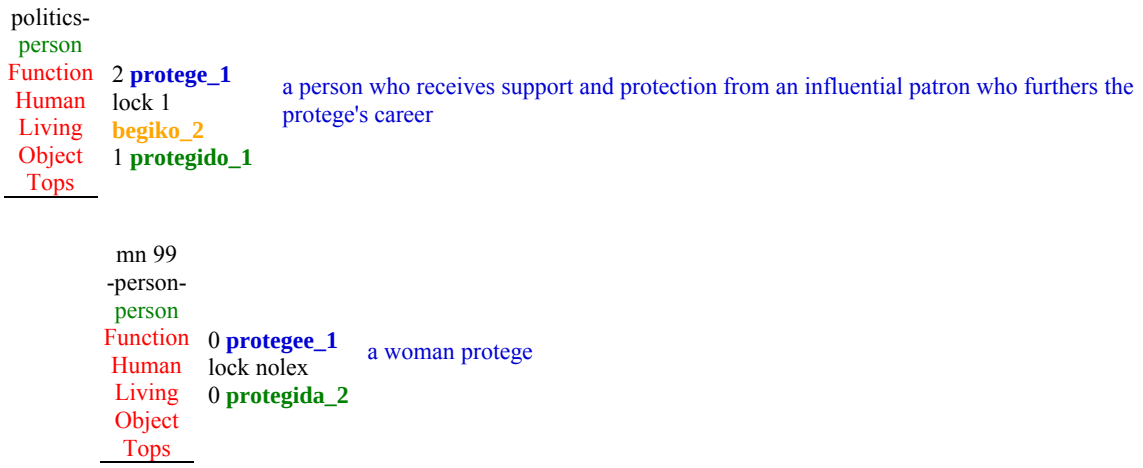
42. irudia

b. Hiztegi edota corpusetan hori egiaztatuko ez balitz, editoreak *partidu politiko* HAULa bakarrik utziko luke.

Bestalde, HAULekin beste irizpide bat izan behar da kontuan, A.3.2.3.1.6 atalean aipatutakoa, hain zuzen ere.

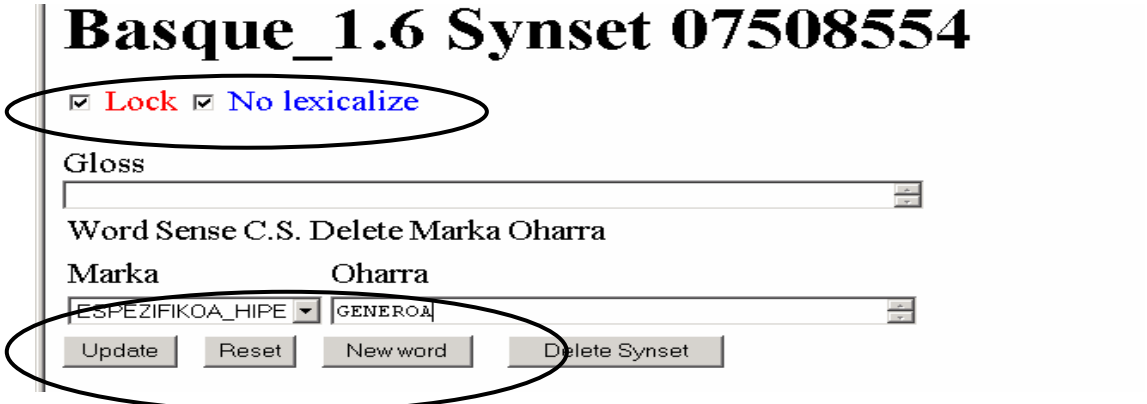
A.3.2.6.7 *Generoa*

Ingeleseko WordNet-en, generoa adierazteko hiponimia erabiltzen dute, hots, hiperonimoa gizonezkoari dagokion synset-a da, eta hiponimoa emakumezkoari dagokiona (ikus 43. irudia).



43. irudia

Euskaraz bi synset-ek ordain ezberdina badute, synset bakoitzean dagokion ordaina gehituko litzateke. Arazoa, ordea, ordaina bera denean dator. Kasu honetan, bi synset-etan *begiko* erabiliko litzateke euskaraz, eta horrelakoetan, emakumezkoari dagokion synset-a ***Espezifikoa Hipe***, ***Nolex*** eta ***Lock*** gisa (ikus A.3.2.3.1.4 atala) markatuko litzateke, eta oharrean *Generoa* idatzi.



44. irudia

Alderantziz gertatuz gero, hots, kontzeptu baten generoa adierazteko ingelesez ordain bakarra izatea (*brother*) eta euskaraz bat baino gehiago (*anaia / neba*), WordNet-en ez dagoen ordain hori *EuskalWordNet-en aurkitu ez diren hitzak* zerrendan apuntatuko da. Hala ere, kasu hau gutxitan geratu(ko) da.

A.3.2.7 Aurrerago lantzekoak

Editorearen eskuliburu hau EuskalWordNet-eko izenak orraztean sortutako zalantzetan oinarrituta dago. Hala ere, zalantza guztiei ezin izan zaie konponbidea aurkitu, eta hurrengo orrazketa baterako utziko dira. Hori egin ahal izateko, editoreak zalantzazkoak diren kasu horiek guztiak aparteko txosten edo zerrendetan gehitzen ditu. Ikusiko dugun bezala, arazo edo zalantza bakoitzari zerrenda bat dagokio[92].

A.3.2.7.1 Kategoria bateraezinak

a) Batzuetan ingeleseko kontzeptu bat euskaratzean, euskaraz beste kategoria bat duela gertatzen da:

Butterfingers (IZE) ⟶ *baldar, dorpe*... (ADJ)

Light (IZE) ⟶ *salir a la luz, kaleratu*, ... (ADI)

Now (IZE) ⟶ *orain* (ADB)

Editoreak horrelako kasuak ***Unlock*** utziko ditu, eta *Kategoria bateraezinak/postposizio* deituriko zerrendan apuntatu. EuskalWordNet-eko adjektiboak, adberbioak eta aditzak lantzean aztertuko dira. Hala ere, horrelako kasuetarako egun pentsatua dagoen proposamena da, kategoria ezberdineko synset-ak **EQ-synonymy** erlazioaren bitartez lotzea:

EusWN				ILI
(ADJ) *baldar, dorpe*	⟶	**EQ-SYN**	⟶	*butterfinger*(IZE)

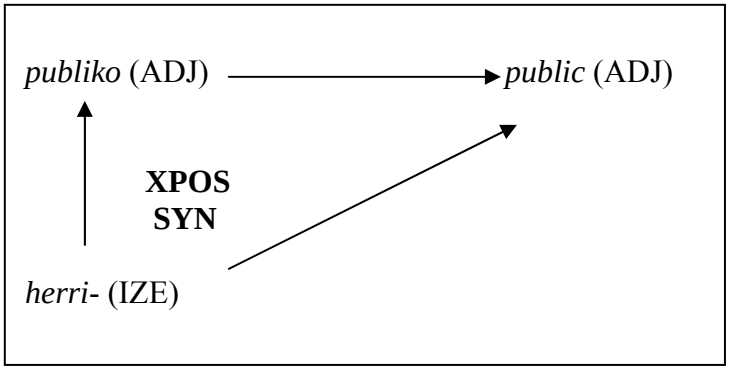
45. irudia

b) Hitz elkarketan: Kasu hauek ere oso bitxiak dira:

- *herri-* (IZE) = *publiko* (ADJ)
- *giza-* (IZE) = *humano*, *social* (ADJ)
- *haur-* (IZE) = *infantil* (ADJ)

Editoreak oraingoz ***Unlock*** utzi, eta dagokien *Kategoria bateraezinak/postposizio* deituriko zerrendan apuntatuko ditu. Aurreko kasuan bezala EuskalWordNet-eko adjektiboak, adberbioak eta aditzak lantzean aztertuko dira. Hala ere, oraingoan ere, aukera bat hurrengoa litzateke:

[92] Berez, eskuliburu honetan azaldu diren erabaki guztiak, horrelako zerrendetatik eratorriak dira, hots, eskuliburu hau sortu arte, editoreak zalantzak guztiak zerrenden bitartez sailkatuak zituen. Beraz, A.3.2.7 atalean agertuko diren kasuak, egunean batean, zerrenda hutsa izatetik eskuliburu honetako irizpide bat izatera pasako dira.



46. irudia

A.3.2.7.2 *Falta diren adierak*

Zenbaitetan editoreak topa ditzake euskaraz eta beste hizkuntzetan lexikalizatuta dauden kontzeptuak, baina ingeleseko WordNet-en lexikalizatuta ez daudenak. Adibidez:

- *liga* = 'txapelketa'
- *kanal* = 'telebista katea'

Hitz hauek EuskalWordNet-en badaude, baina **adiera zehatz horrekin ez**. Oraingoz, editoreak horrelakoak *EuskalWordNet-en aurkitu ez diren adierak* zerrendan jasoko ditu, eta geroago hauek EuskalWordNet-en sartzen hasteko asmoa baitago.

Bestalde, horrelako arazoen aurrean, ingeleseko WordNet 2.0 bertsioa kontsultatzea komenigarria da, 1.6 bertsiotik 2.0 bertsiora adiera berri hori txertatua egon daitekeelako.

A.3.2.7.3 *Kontzeptu kulturalak*

Atal honetan kontzeptu *kulturalak* deritzogunak sartzen dira. Kasu honetan ingeleseko WordNet-en ez dauden adierak dira, euskal kulturarekin loturik daudelako. Aurreko atalean ez bezala, **hitz hauek ez daude** WordNet-en, ingelesez kontzeptu horiek ez direlako existitzen, hain zuzen ere. Adibidez: *pilotari, bertsolaritza, euskaldun, kalimotxo, sagardotegi, txapela, euro* eta abar.

Honelako kasuen aurrean, momentuz editoreak dagokien zerrendan (*EuskalWordNet-en aurkitu ez diren hitzak*) jarriko ditu aurrerago lantzeko asmoz.

A.3.2.7.4 *Posposizioak*

Editorea posposizio baten aurrean aurkitzen denean, momentuz *Kategoria bateraezinak/postposizio* zerrendan jarriko du, adjektibo, adberbio eta aditzetara iristean landuko baita. Adibidez:

ondo ⟶ *-re ondora/ondoan*

albo ⟶ *-re albora/alboan*

A.3.2.7.5 *behar, uste, ahal... bezalako formak*

Aditz perifrastikoen kasuan ere editoreak momentuz *Hutsak* bezala markatuko ditu `~jirhizts/Corpus/PROFIT2/koordinazioa` katalogoan (honen berri 4.1 atalean eman da); hauek hurrengo fase batean landuko dira.

A.3.2.7.6 *Unlock uzten direnak*

Irizpide hauekin nahikoa ez bada eta synset bat ***Unlock*** utzi nahi bada, synset hori zalantza-zerrenda batean apuntatu egin behar da, zalantzaren zergatiarekin batera, gero lan-taldearekin komentatzeko eta zalantza mota horri konponbideren bat topatzeko. Hala ere, hau gutxitan gertatu behar da eta gertatuko balitz, garrantzizkoa da ***Unlock*** uzten ditugunak, zalantza garrantzitsuenak izatea.

A.4 Ondorioak

Lan honen helburu nagusiena eskuliburu bat sortzea izan denez, kezkarik nagusiena ulergarria eta erabilgarria gertatzea izan. Abiapuntua honako hau izan da: editore lanetan aritzeko hizkuntzalari hasi berriarentzako informazioa jasotzea; besteak beste, interfaze desberdinen erabilera, beharrezko tresna guztien argibideak, orrazketarako irizpideak,... jasotzen ditu.

Bestalde, eskuliburu hau ez da hemen itxita geratzen. Eguneratuz joango den zerbait da, aurreko erabakiak berritu eta sortu berriak txertatu beharko dira. Honela, ondoren datorren taulan adierazi nahi dugu zein erabaki diren finko edo zein dauden oraindik eztabaidapean, eta dagoeneko zeintzuk aplikatzen diren eta zeintzuk ez.

IRIZPIDE MOTAK	IRIZPIDEAK	FINKOAK BAI	FINKOAK EZ	APLIKATZEN DIRA	EZ DIRA APLIKATZEN
SYNSET mailakoak (NOLEX)	*Nolex arrunta*	X		X	
	Espezifikoa Nolex	X		X	
	Orokorra Nolex	X		X	
	Espezifikoa Hipe (Nolex)	X		X	
	-TU / -T(z)E	X		X	
	Bestelako kasuak	X		X	
VARIANT mailakoak	*RARE*	X		X	
	PLU	X		X	
	IXALEX	X			
IDAZKERA arazoak	*Marratxoak*	X		X	
	Artikulua daramatenak	X		X	
	HAULak idazteko era desberdinak	X		X	
	Hizki larriak eta xeheak	X		X	
BESTELAKOAK	*-keta, -kuntza, -mendu bezalakoak*	X		X	
	Hiztegiak bat ez datozenean	X		X	
	Antzeko synset-ak	X		X	
	Adieren egokitasuna	X		X	
	Figuratiboak	X		X	
	Generoa	X		X	
AURRERAGO lantzeko	*Kategoria bateraezinak*	X			X
	Falta diren adierak	X			X
	Kontzeptu kulturalak	X			X
	Postposizioak		X		X
	Aditz Perifrastikoak		X		X
	Unlock uzten direnak		X		X

A.5 Hiztegi terminologikoa

datu-base lexikala
Lexikoari buruzko informazioa modu egituratuan gordetzeko euskarri informatikoa

Euskararen Datu-Base Lexikala (EDBL)
Euskararen tratamendu automatikorako euskarri lexikal orokorra.

Ezagutza-base lexikalak (EBL)
Hitz eta adierei buruzko informazioa duten lexikoiak. EBLen ezaugarri garrantzitsuena herentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen dira-eta.

hiperonimo
Hitz bat bestearekiko orokorragoa denean.

hiponimo
Hitz bat bestearekiko zehatzagoa denean.

interfaze
Konputagailuko bi osagaien arteko komunikazioa ahalbidetzen duen bitarteko fisiko edo logikoa.

synset
Kontzeptu bat adierazteko erabiltzen den sinonimoen multzoa

variant
Synset bakoitzean hizkuntza bakoitzeko dagoen ordaina.

Nolex
Adiera bat euskaraz lexikalizaturik ez dagoela adierazteko termino edo marka da

B Synset-en glosak itzultzeko irizpideak

Ikusi dugu (ikus A.2 atalean), EuskalWordNet-en interfazea aztertu dugunean, synset-ek badutela GLOSA bat: adiera ulertzeko azalpen/definizio bat. Bada, etiketatzailearen azken zereginetako bat hitzaren synset-en glosak itzultzea da. Behin hitza etiketatzen bukatu duenean, etiketatzailea glosa itzultzeko prest dago, berak landu baititu hitzaren agerpen desberdinak.

Itzulpenerako zenbait irizpide bildu dira, idazkera, egitura, estiloa,... eta beste kontuan harturik. Hona hemen irizpideak.

B.1 Glosak: irizpide orokor batzuk

a. Lehenik eta behin, hiztegietan (*Hiztegixa*, *Harluxet*....) definizio aproposik badagoen begiratu. Hala bada, erabili edota moldatu.

b. Metahizkuntza ez erabili, adibidez: DIRU "ondasun handia adierazteko hitza" ("adierazteko hitza" bezalakoak ekidin).

c. Ingeleseko glosetan bigarren pertsona ageri den kasuetan, euskaraz inpertsonala erabiliko dugu.

Adibidea: *lan*[93]

base concept		
act		
Agentive	129 **employment_2 job_1 work_3**	the occupation for which you are paid
Cause	lock 61 **lan_11 enplegu_2 lanpostu_1**	soldataren truke egiten den jarduera profesionala; "lan bila dabil"
Dynamic	249 **business_6 job_1 line_19 line_of_work_1**	the principal activity in your life that you do to earn money
Purpose	**occupation_1**	Actividad que realiza una persona a cambio de un salario
Social	lock 131 **trabajo_2**	
Static		
UnboundedEvent		

1go irudia

d. Genusak: Zeintzuk dira euskaraz erabili behar diren izen orokorrak?
   - **pertsona**: gauza psikologikoez ari garenean.
   - **gizaki**: orokorragoa da.
   - **Genus** bat erabiltzea zaila gertatzen bazaigu, erlatibozko perpausaren bidez adierazi. Adjektiboak definitzeko erlatibozkoa erabili beharko da.
   - "**zera**" komodina ahal dela ez erabili. Adibideak: *gizaki, pertsona multzo, duen pertsona, duena...* Testuinguruak lagunduko du genusa aukeratzen.

[93] Adibideetako irudietan ikus daitekeen bésala, EuskalWordNet 1.6, Gaztelaniako WordNet 1.6, eta ingeleseko WordNet 1.6 (urdin argia) eta 1.7 (urdin iluna) agertzen dira. Ingeleseko 1.7 bertsioak definizio landuagoak izan ditzake, eta kontsulta egitean komenigarria da hau ere egotea.

Adibidea: *herri*

common_people_1	people in general
folk_1	biztanleen gehiengoa osatzen duen gizaki multzoa, belaunaldiz belaunaldi bertako kultura
lock 2 **herri_1 populu_2**	eta ohiturak gordetzen dituena; "herriak markatzen du talde-izaera, eta hark gordetzen ditu
5 **common_people_1**	ohiturak belaunaldiz belaunaldi"
folk_1	people in general
lock 3 **plebe_1 vulgo_1**	Grupo de gente que constituye la mayoría de la población y que define y mantiene la
pueblo_1	cultura popular y las tradiciones

base	11 **body_politic_1 commonwealth_2 country_2**	a politically organized body of people under a
concept	**land_9 nation_1 res_publica_1 state_3**	single government
group	lock 4 **herri_4 estatu_3 nazio_2**	unitate politiko bakarraren baitan dagoen gizaki
Function	11 **body_politic_1 commonwealth_2 country_2**	multzoa; "herriak agintaria hautatu du"
Group	**land_9 nation_1 res_publica_1 state_3**	a politically organized body of people under a
Human	lock 11 **estado_2 país_1**	single government
		Grupo de gente regida por un único gobierno

2. irudia

e. Definizioetan ez dugu landu beharreko hitza erabiliko, ezta adiera bakoitzari dagokion aldagaietako (variant-etako) bat ere.

Adibidea: *diru*

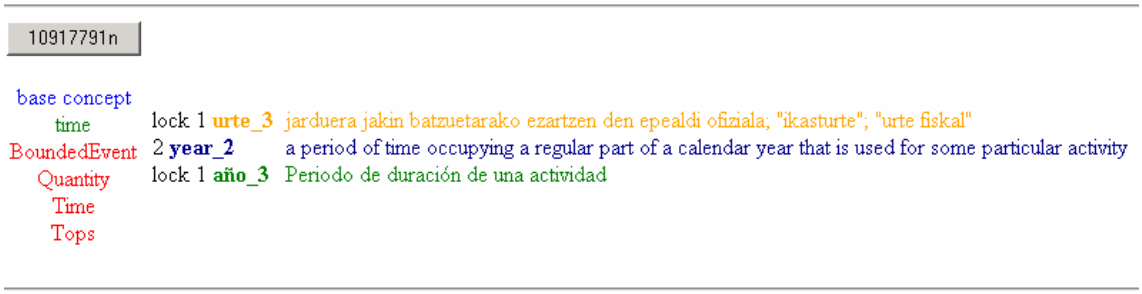
base concept		wealth reckoned in terms of money
possession	2 **money_2**	ondasun handia adierazteko hitza; “diru#familia horrek dirua du”
Artifact	lock 1 **diru_2**	wealth reckoned in terms of money
Function	2 **money_2**	
MoneyRepresentation	lock 2 **dinero_1**	

3. irudia

f. Ahalik eta aditz gutxien erabili.

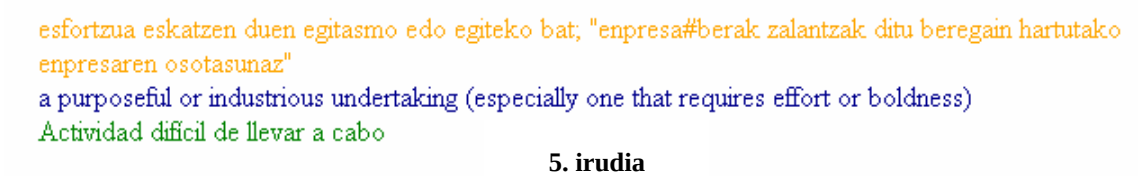
B.2 Egitura

Lehenik definizioa emango da. Adibideak gehituz gero, definizioaren amaieran puntu eta koma jarri, hutsunea utzi ondoren, eta adibideak komatxo artean gero. Adibide bat baino gehiago jarriz gero, puntu eta komaz bereiziko dira: _____(def)_____; “_____”; “_____”



4. irudia

Adibideak, esan bezala, komatxo artean joango dira; definizioa eman baino lehen bertan agertuko den variant-a zehaztuko da, eta ondoren “almohadilla” delako ikurra.



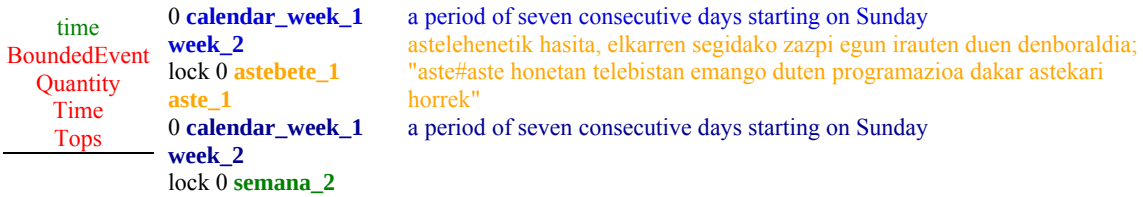
5. irudia

B.3 Arazoak eta erabakiak

a. Interferentzia kulturalak direla medio, hitzen definizioetan ematen den mundu-ikuspegia zenbaitetan ez dator bat euskaraz eta ingelesez.

ERABAKIA: Euskarako definizioa ingeleseko glosan oinarrituta emango dugu. Ingeleseko oinarri hori euskararako baliagarri ez den kasuetan, euskal hiztegi elebakarretara joko dugu (*Euskal Hiztegia*, *Hiztegi Batua*, *Euskal Hiztegi Modernoa*)

Adibidez: *aste*



6. irudia

Ingelesez igandea da astearen lehen eguna, euskaraz, aldiz, astelehena.

b. Etiketatzen hasi baino lehen, hitz bakoitzak euskaraz zein adiera har dezakeen aztertzen du lan-taldeak. Bilera hauetan erabakitzen dugu hitzaren adiera (synset) bakoitzean zein testuinguru sartuko dugun. Zenbaitetan ingeleseko definizioa eta guk erabakitakoa ez datoz bat.

ERABAKIA: Honelakoetan, hiztegietako definizioetara jo. Adibidez, *defentsa*: ingelesez defentsa hitzaren adieretako bat futbol amerikarrean funtzio zehatza duen jokalariarena da. Gurean kirol horrek jarraitzailerik ez duenez, hutsik utzi, eta arruntagoa gertatzen zaigun futboleko defentsa beste adiera orokorrago horretan sartuko dugu[94].

person		
Function	7 **back_4**	a person who plays in the backfield
Human	lock 3 **atzelari_2**	futbolean atzeko postuan jokatzen duena; "defentsa#ohiko taktika defentsiboa
Living	**defentsa_9**	zelairatuko du: bost defentsa, hiru erdilari, erdi-punta bat eta aurrelari bat"
Object	8 **back_4**	a person who plays in the backfield
Tops	lock 6 **defensa_9**	Jugador que evita que la pelota del equipo contrario llegue al área

act	0 **field_general_2 quarterback_2**	
Agentive	**signal_caller_2**	the position of the football player in the backfield who directs the offensive play of his team
Cause	lock 0 **defentsa_11**	
Dynamic	0 **field_general_2 quarterback_2**	
Purpose	**signal_caller_2**	the position of the football player in the backfield who directs the offensive play of his team
	nolex 0	

7. irudia

c. Irizpide orokorretan (ikus 1.1.5. puntua) lantzen ari garen hitza ez dugula erabili behar esan badugu ere, WordNet-en izaera dela eta, batzuetan honelakoak ekiditea ezinezkoa gerta liteke. Zenbait kasutan ezinbestean erabili beharko da, dena dela, hitz batek duen erabilerarik arruntenean ez erabiltzen saiatu behar gara.

Adibidez: *aldizkari*

group	0 **magazine_3**	
Function	**magazine_publisher_1**	a business firm that publishes magazines
Group	lock 0 **aldizkari_7**	aldizkariak argitaratzen dituen enpresa; “aldizkari# aldizkari batean lan egin nahi nuen”
Human	0 **magazine_3**	
Occupation	**magazine_publisher_1**	a business firm that publishes magazines
	nolex 0	

8. irudia

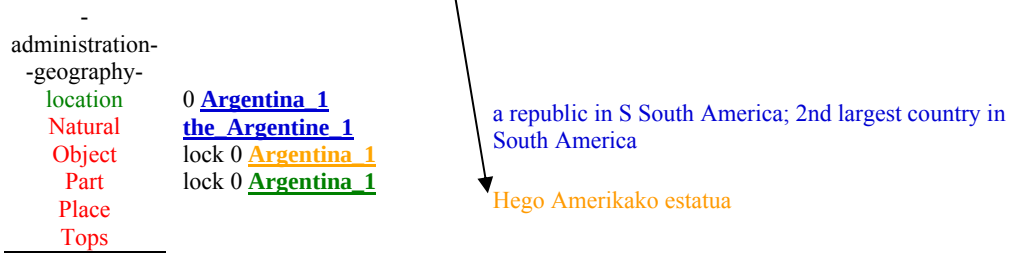
d. Herrialdeen synset-ak (*Argentina, Peru, Luxenburgo...*) definitzean, *Harluxet Hiztegi Entziklopedikoan* dagoen lehenengo esaldia hartuko da.

Adibidez: *Argentina*

[94] Adibide honetan, komenigarria da editoreari komentatzea, behar bada, *defentsa* kontzeptu orokorrago hori EuroWordNet-en egon daiteke eta editorea ez da konturatu, edota EuroWordNet-en ez dago eta sortzea erabakitzen da.

Argentina

(República Argentina). Hego Amerikako estatua. 23 probintzia eta Barruti Federal batek osatzen dute. Mugak: I-an Bolivia eta Paraguai; E-an Brasil, Uruguai eta Ozeano Atlantikoa; H-an eta M-an Txile. Erliebea bi unitate handiz osatuta dago: Andeak M-an eta lautadak E-an...



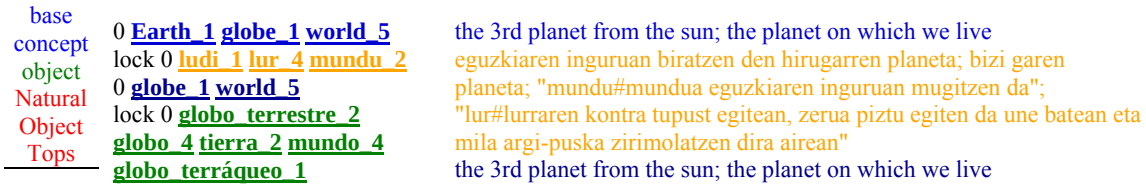
9. irudia

B.4 Puntuazioa: irizpide orokorrak

a. Puntu eta koma ekidin definizioetan. Hitz baten definizioan bi azalpen desberdin agertzen direnean soilik erabiliko dugu puntuazio ikur hau. Definizio batek kontzeptu bakarra badu, puntu eta komarik ez erabili.

b. Izen bereziak desberdintzeko asmoz, ez dugu letra larririk erabiliko izen arruntetan, ez definizioetan, ezta adibideetan ere. Glosaren amaieran, gainera, ez dugu punturik jarriko.

Adibidez: *lur*



10. irudia

c. Zenbaki ordinalak letraz idatzi (ikus aurreko adibidea)

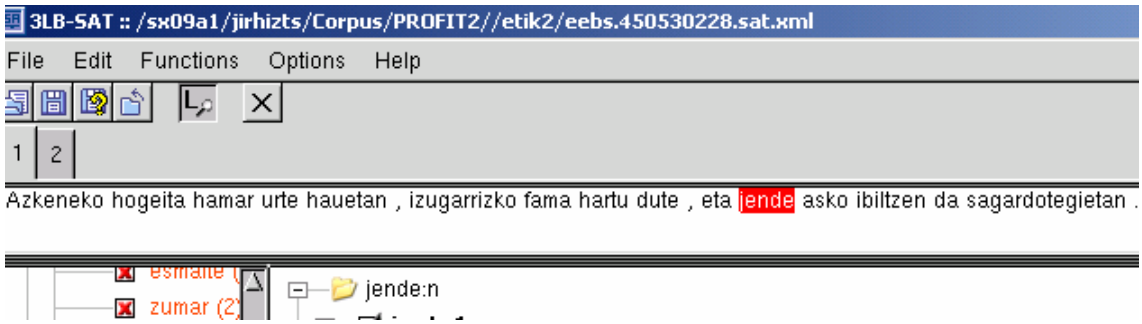
B.5 Adibideak

a. Adibideak gehitzeko, etiketatzeko baliatzen dugun corpusera joko dugu lehenik. Bertan adibiderik aurkituko ez bagenu, edo hauek egokiak izango ez balira, orduan ingeleseko WordNet 1.6-eko Browser-ean bilatuko genuke[95]. Browser-a baliagarri ez den kasuetan, *XX. mendeko Euskararen Corpus Estatistikoa, Ereduzko prosa* gaur, hiztegietako adibideak, bestelako corpusak... Baliteke itzultzen ari garen adiera horretarako etiketatzen ari garen hitza euskaraz gehiegi ez erabiltzea; kasu horietan ez da adibiderik ezarri beharko.

b. Etiketatzeko baliatzen dugun corpuseko esaldiak gehienetan luzeegiak izan ohi dira. Honelakoetan, hartzen dugun adibideak mendeko perpausa badu, eta luzeegia gertatzen bada, hau egokitu mendekoa ezabatuz. Edota lantzen ari garen hitza mendeko esaldi batean badago, esaldia nagusi bihurtuko dugu.

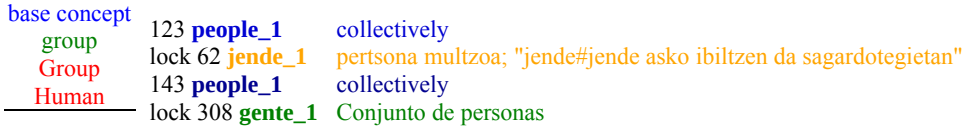
   Adibidez: *jende*

   Corpusean:



11. irudia

   Glosan:



12. irudia

c. Ezin dugu euskaraz arruntagoa gertatzen den hitz bat bilatu eta ondoren guri interesatzen zaigun hitzaren (sinonimoaren) ordez jarri. Euskaraz definitzen duguna adierazteko hitz bat erabiltzen ez bada, inon ez badugu adibiderik topatzen, ezin dugu beste hitz batekin erabiliko genukeen adibidea moldatu.

   Adibidez: *abendu* ('miru' adierarekin ez da erabiltzen, beraz, adibiderik ez)

[95] Browser-a erabiltzeko azalpenak A.3.1.1 atalean daude.

HONAKO HAU EZIN DA EGIN:

asmo (‘amarru’ adierarekin)

Ereduzko Prosa Gaur:
Bilaketa egiteko *amarru* erabili dugu nahiz eta gero *asmo* agertuko den.

Friedelek lehen gutun hura besterik ez zuen idatzi; hura guztia **amarru** itsusi bat zela adierazi zuen, eta ez zuela hartan zerikusirik gehiago izango.

	3 **contrivance_3**	
	dodge_1 stratagem_2	an elaborate or deceitful scheme contrived to deceive or evade
cognition	lock 2 **asmo_5 amarru_4**	ihes edo irizur egiteko egiten den gezurrezko plana; "asmo#hura guztia
3rdOrderEntity	**azpikeria_6**	asmo itsusi bat zela adierazi zuen, eta ez zuela hartan zerikusirik
	3 **contrivance_3**	gehiago izango"
	dodge_1 stratagem_2	an elaborate or deceitful scheme contrived to deceive or evade
	lock 4 **estratagema_3**	

13. irudia

d. Bi hitzez osaturiko adibideak erabil daitezke:

	2 **field_10**	a set of elements such that addition and multiplication are commutative and associative and multiplication is distributive over addition and there are two elements 0 and 1
group	lock 0	honako propietateak dituen elementu multzoa: trukakorra, elkarkorra eta banakorra;
Tops	**gorputz_7**	“gorputz#gorputz trukakorra”
	2 **field_10**	a set of elements such that addition and multiplication are commutative and associative and
	nolex 1	multiplication is distributive over addition and there are two elements 0 and 1

14. irudia

C `landutakoak.n.usuenak.txt` eta `landutakoak.n.ezusuak.txt` fitxategietako markak

C.1 *Editoreak jartzen dituen markak*

- **ERR** (lematizazio errorea)
- **IZB** (EuskalWordNet-en eta ingeles/gaztelaniako WordNet-etan ez dagoen izen berezia, oraingoz ez landu)
- **HUTSA** (postposizio edo aditz perifrastikoak: *ahal, behar, aurre, alde...*)
- **ZKI** (digituak)
- **EWN** (hitza ez dago EuskalWordNet-en, ezta ingeles/gaztelaniako WordNet-etan ere : *sagardotegi, euskal...*)
- **???** (arazoak dituen hitza, oraingoz langu gabe utziko dena)
- **etik** (hitza etiketatzen ari dira)
- **etikOK** (hitza etiketatzen amaitu dute, eta epaitzeko prest dago)

C.2 *Epaileak jartzen dituen markak*

- **epai** (hitza epaitzen ari dira)
- **epaiOK** (hitza epaitzen amaitu da)
- **epaiOKC1S** (hitza epaitzen amaitu da)

C.3 *Etiketatzaileek jartzen dituzten markak*

- **AUTO** (monosemikoak; automatikoki etiketatuko direnak)
- **AUTOC1S** (polisemikoak, editoreak landu behar dituen hitzak)