EUSKARA-GAZTELANIA TERMINOLOGIA ELEBIDUNAREN ERAUZLE AUTOMATIKOA

Itzulpen-memorietatik terminologia elebiduna automatikoki erauzteko prototipoa

Antton Gurrutxaga, Xabier Saralegi, Sahats Ugartetxea

Elhuyar Fundazioa

Iñaki Alegria

IXA taldea

1 Sarrera

Artikulu honetan, Elhuyar Fundazioak eta IXA taldeak terminologia-erauzketa elebidunaren alorrean egindako ikerkuntza- eta garapen-lanak aurkeztuko ditugu. Proiektu honen helburuak da es-eu itzulpen-memorietatik termino baliokideen bikoteak automatikoki erauzteko teknikak lantzea eta teknika horiek inplementatuz prototipo bat garatzea. Helburu hori ikertze-alor zabalago baten baitan kokatuta dago. Izan ere, itzulpen-memoriak corpus eleaniztunen eta, zehatzago, corpus paraleloen kasu partikular bat dira; bestetik, terminoa hizkuntza-unitate mota bat da, azpimultzo bat. Beraz, bi hizkuntzatako termino baliokideak erauztea baliokidetza lexikalen erauzketaren alorreko kasu partikular bat da.

2 Motibazioa eta interesa

Informazioaren Gizartean eleaniztasunak duen garrantzia kontuan harturik, bereziki nabarmentzekoak dira testu edo corpus paraleloen ustiaketak eskain ditzakeen aukera berriak. Corpus paraleloak interes handikoak dira, hizkuntzen arteko baliokidetzei buruzko informazio asko atera daitekeelako bertatik. Informazioa alor askotan da baliagarria: hiztegigintza elebiduna, terminologia, itzulpen automatikoa, itzulpengintzari laguntzeko tresnak, itzulpengintzaren teoria eta praktika, hizkuntza-irakaskuntza, hizkuntzalaritza kontrastiboa...

Aplikazio askotan behar diren baliabide lexikal eleaniztunak edo, gehienetan,

elebidunak eskuratzeko, lehen aukera da erabiltzea argitaraturik dauden edo eskuragarri diren hiztegi elebidunak, datu-base terminologikoak, ezagutza-base eleaniztunak (Euskal WordNet) eta abar. Horrek muga batzuk ditu, ordea. Hain zuzen ere, hainbat autorek erakutsi du hiztegietan biltzen den informazio lexikal eleaniztuna eta testuetan dagoena ez datozela beti bat, batez ere, hizkuntza azkar garatzen eta aberasten delako, eta testuetako baliokidetza lexikal asko oraindik ere hiztegietara heldu ez direlako, batik bat espezializatuak badira.

Horrek guztiak baliabideak testuetatik bertatik erauzteko beharra jartzen du agerian. Hizkuntza Teknologien bidez, informazio elebidun hori erauzteko teknikak garatzeko aukera dago. Hizkuntza Teknologien beste hainbat alorretan bezala, honetan ere euskararen izaerak, hau da, hizkuntza eranskaria izateak, teknika linguistikoak erabiltzera behartu gaitu (teknika estatistikoekin konbinatuta, jakina), testu gordinaren gaineko teknika estatistiko hutsen emaitzak barreiatuak izaten direlako.

3 Corpus eleaniztunak: definizioa eta motak

Corpus eleaniztunak honela defini litezke: *Zenbait parametroren arabera ezaugarri komunak agertzen dituzten bi hizkuntzatako edo gehiagotako testu-bildumak*.

Bi corpus eleaniztun mota nagusi daude:

- *Corpus paraleloa*: eduki 'bera' duten testuez osatuak, ia beti jatorrizko testu batez eta beste hizkuntza batera edo batzuetara egindako itzulpenez osatutako testu-multzoa da. *Bitestu* ere esaten zaie
- *Corpus konparagarria*: corpus eleaniztuna osatzen duten testuak bata bestearen itzulpen ez direnean, baina hainbat ezaugarri komun dituztenean (eremua, generoa, denbora-bitartea, komunikazio-helburua...)

Corpus paraleloen kasu bakunena, eta ohikoena, testu-bikoteez osatutako corpus elebiduna da. Corpusa osatzen duten testu-osagaien artean baliokidetzak ezartzen direnean, corpus paralelo 'parekatuak' direla esaten da.

4 Corpus paraleloak

4.1 Parekatzea

Corpus paraleloak hizkuntza bakoitzeko testuen artean dagoen parekatze-mailaren

arabera bereizten dira. Parekatzea (*alignment*) honela defini daiteke:

> "Aligning consist in finding correspondences, in bilingual parallel corpora, between textual segments that are translation equivalents." (Kraif, 2002b)
>
> "By 'align' is meant the association of chunks of text in the one document with their translation or equivalent text in the other document". (Somers)

Parekatzearen kontzeptuan sakontzeko, lehen egitekoa itzulpen-baliokidetza bera zer den argitzea da. Hein handian, autore gehienak ados daude baliokidetza ezaugarri globala dela, funtzionala, hau da, maila pragmatikoan edo 'komunikatiboan' ezarri behar dela. Dena den, ikuspegi horretatik ezar litekeen lehen baliokidetza-maila testu-mailakoa bera izan liteke. Baina, egia izanik ere, baieztapen hori ez da oso praktikoa, eta gehiago zehatu edo 'zatikatu' beharra aitortu ohi da. Zatikatze horren gakoa 'itzulpen-unitatea' da (*translation unit*).

Itzulpen-unitatea zer den eta unitateok nola sailka daitezkeen aztertzerakoan, Joseba Abaituak egindako lanean oinarritu gara (Abaitua, 1997). Hemen, oinarrizko ideiak baizik ez ditugu aditzera emango. Vinay eta Darbelnet-ek honela definitu zuten itzulpen-unitatea:

> "The smallest segment of the utterance where the cohesion of sings is that they cannot be translated separatelly." (Vinay *et al.*, 1958)

Nolanahi ere, itzulpen-baliokidetzaren kontzeptua erlatiboa da. Adibidez, baliokideak diren testu- edo esaldi-unitateen barnean baliokide 'atomikoagoak' ere egon daitezke. Hona hemen Abaituak baliokidetzat duen itzulpen-unitatea, 'unidad de traducción fraseológica' delakoa definitzean (Abaitua, 1997):

<UTF_1>Mediante la Orden Foral de referencia se ha dispuesto lo siguiente:

<UTF_1>Aipameneko Foru Aginduaren bidez honako hau xedatu da

Bistan da hor esalditik beherako baliokidetzak ere ezar daitezkeela:

Orden Foral ⇔ Foru Agindua

Se ha dispuesto ⇔ xedatu da

Erlatibotasun horren baitan, aipatu adibideotako bi ikuspegiak izan daitezke egokiak, aztergaia, proiektuaren helburua edo aplikazioa zein diren.

4.2 Parekatze-mailak

Lehenago adierazi dugu corpus paraleloak parekatze-mailaren arabera bereizten direla. Hauek dira parekatze-maila ohikoenak:

- Dokumentuak: erakunde, enpresa eta abarren dokumentu-biltegi egituratuak; Web gune eleaniztunak...
- Esaldiak (1:n): itzulpen-memoriak (TMX), TEI P4 ereduaren arabera parekatuak... Dokumentu-bikote elebidunak automatikoki parekatzeko tresnak badaude, eta aipatzekoa da, zer esanik ez, euskararako berariazko sistema garatzeko egin den lana (Casillas *et al.*, 2004)
- Esalditik beherako itzulpen-unitateak
  - Unitate lexikalak
    - Hitz bakunen artekoak (1:1) (*one-to-one alignment*)
    - Unitate lexikalen artekoak (n:m): hitzak, terminoak, lokuzioak, kolokazioak...
  - Entitateak (izen bereziak; datak)
  - Bestelako diskurtso-unitateak: formulismoak, itzulpen-txantiloiak...

5 Corpus paraleloen ustiaketa

5.1 Helburuak: parekatzea vs erauzketa

Hizkuntza Teknologien ikuspegitik, hau da helburu interesgarriena: *corpus paraleloan dagoen informazio inplizitua ustiatzea/esplizitatzea*. Horren aplikazio nagusiak dira baliabide lexikalak eratzea (hainbat zereginetarako behar direnak: hiztegigintza, ezagutza-base eleaniztunak, informazio-erauzketa eleaniztuna...) eta itzulpen automatiko zein lagundutako sistemak hobetzea.

Ikusi dugunez, esplizitatze-lan hori baliokidetzak ezarriz egiten da, eta maila askotakoa izan daiteke. Proiektu honetan, esalditik beherako unitateen arteko baliokidetzak izan ditugu jomuga. Helburu horrekin garatu diren sistemak, gehienetan, esaldi-mailan parekatutako corpus paraleloetatik abiatzen dira, eta bi emaitza-mota izan ohi dituzte:

- Hitzez hitzeko parekatzea (*full word alignment*)
- Baliokidetza lexikalak erauztea: lexikoi elebidunak

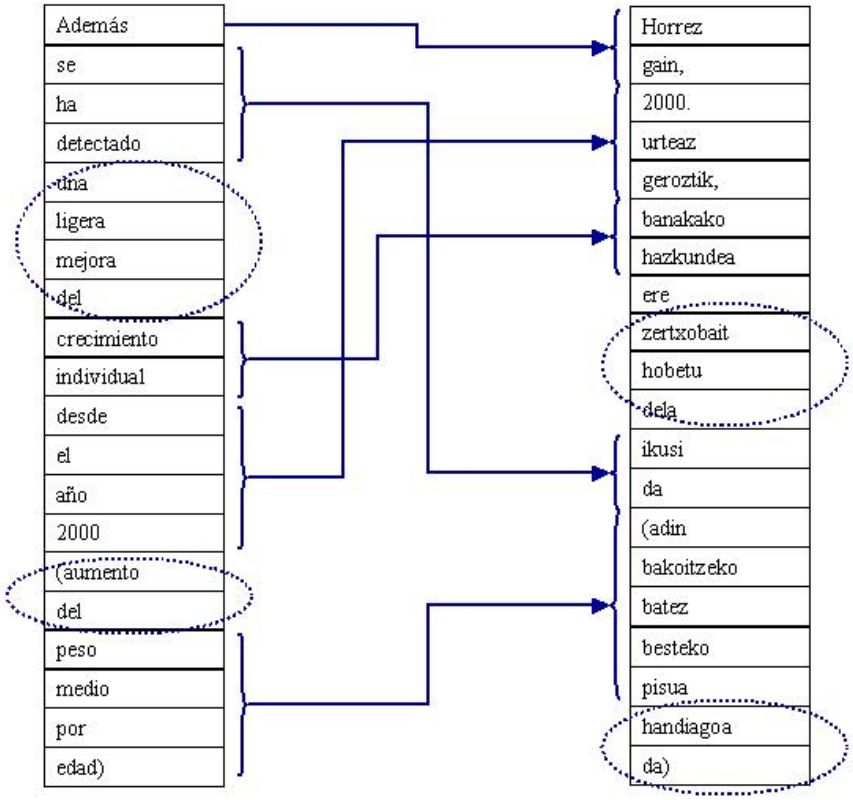
Hitzez hitzeko parekatzea egiten denean, helburua da jatorriko testuko eta itzulpeneko osagaien artean itzulpen-baliokidetzan oinarritutako parekatze 'osoa' egitea (*full word alingment, high resolution alignment*). Beraz, esalditik beherako unitatetan zatikatzen dira testuak, itzulpen-baliokidetzan oinarrituta. Horrek esan nahi du, adibidez, itzuli den zerbait ezin dela parekatu gabe utzi, eta zerbait itzuli ez denean 'null' elementua esleituko zaiola itzulpen-testuan. Parekatzen diren esaldi-atalak, ikuspegi linguistikotik, era askotakoak izan daitezke (perpaus nagusiak menpekoak, sintagma osoak, sintagma-atalak, esapideak...). Gainera, ezartzen diren parekatzeak itzulpenaren testuinguruan dira esanguratsuak, eta, itzulpen-estrategiak desberdinak izan daitezkeenez, batzuetan, baliokidetzak itzulpen jakin baten testuingurura mugatuta egon daitezke.

Beste ikuspegia corpus paralelotik baliokidetza lexikalak 'erauztea' da. Oraingoan, kontzeptu giltzarria baliokidetza lexikala da (*lexical correspondence*). Honela definitu da:

> "A relation of denotational (conceptual, extra-linguistic) equivalence between two lexical units in the context of two segments that are translation equivalents." (Kraif, 2002b)

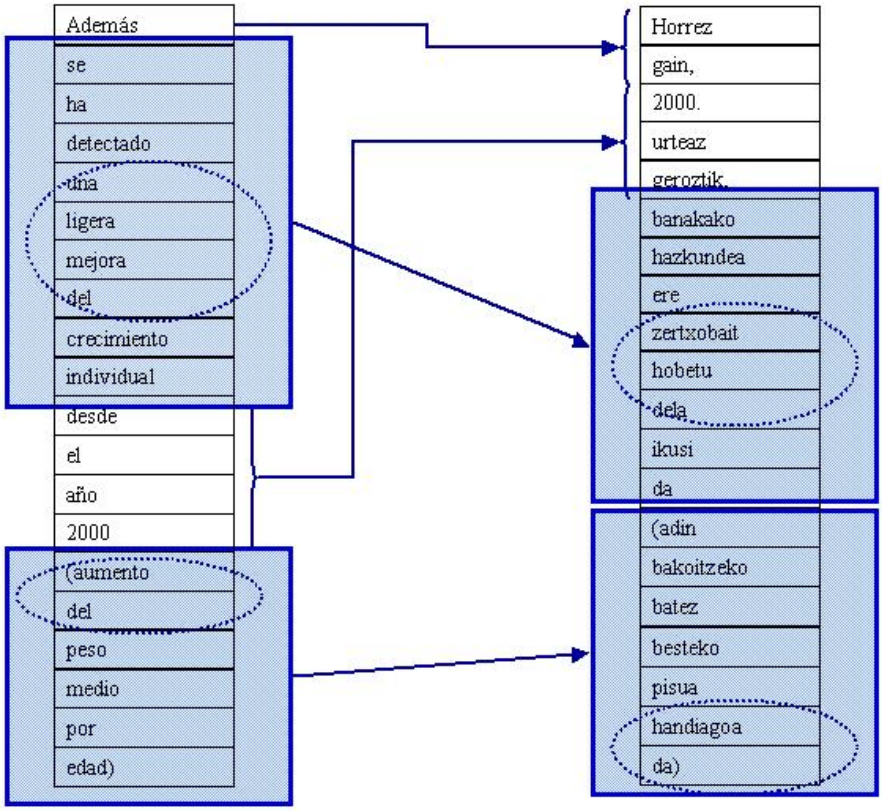
Baliokidetza horiek hizkuntza-unitateen artekoak dira, eta, beraz, segmentazioa egiteko irizpidea elebakarra da lehendabizi (Kraif, 2002b). Hizkuntza-unitate lexikalen arteko baliokidetzak maila denotazionalean ezartzen dira, eta, berez, testuinguru anitzetan dira baliagarri, ez soilik prozesatzen ari den esaldi-bikoteen testuinguruan. Horrek guztiak berekin dakar estrategia honetan ez dela nahitaezkoa testuko elementu guztiak parekatzea (Merkel & Ahrenberg, 1999; Kraif, 2002a; Tiedemann, 2003).

Hurrengo adibidean, argiago ikus daiteke hori:



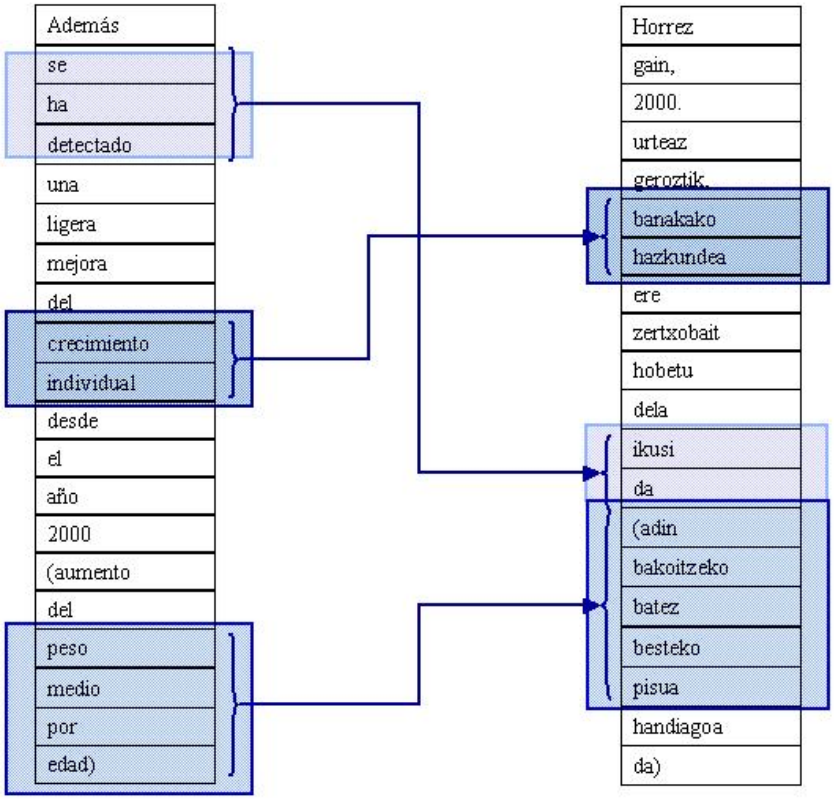
1. irudia. Itzulpen-unitateen ezin parekatuzko zatiak (birbilen barnekoak).

Gezien bidez loturik ageri dira ezar daitezkeen esalditik beherako itzulpen-unitateak. Biribilez inguratuak, ordea, ezin dira parekatu. Horren arrazoi nagusia da ez direla itzulpen-unitate 'atomikoak', hau da, itzulpen-unitate handiago batzuen zatiak direla, itzulpen-baliokidetza horien gaineko esaldi-atalen artean gertatzen baita. Hurrengo irudian, urdindurik ageri dira itzulpen-unitate horiek.



2. irudia. Erabateko parekatzea.

Beraz, erabateko parekatzea egin ahal izateko, horiek dira ezarri beharreko itzulpen-baliokidetzak. Bistan da itzulpen-unitate horietako batzuetan badirela unitate lexikalen arteko baliokidetzak. Baliabide lexikalen erauzketaren ikuspegitik, berriz, horiexek dira interesgarriak. Terminologiara bideratutako erauzketa batean, honakook lirateke xedea:



3. irudia. Terminologia-unitateen erauzketa.

Proiektu honetan, baliokidetza lexikalak erauztearen bidetik heldu diogu corpus paraleloen ustiaketari. Esan gabe doa, ustiatze-modu bata zein bestea dira interesgarriak, aplikazioa zein den.[1] Baliabideen erabilgarritasuna eta aplikazio-eremuen aniztasuna gogoan izanik, erauzketa-tresna egitea lehenetsi dugu.

5.2 Metodologiak

Hitzez hitzeko parekatzerako zein baliokidetza lexikalak erauzteko egin diren saiakuntzak eta garatu diren sistemak bi metodo nagusitan oinarrituta daude (Hiemstra, 1999: 15; Tiedemann, 2003: 12):

- Elkartze-neurrietan oinarrituak
- Estimazioan oinarrituak (itzulpen-eredu probabilistikoetan)

Bi eredu horiek teknika estatistikoetan oinarrituta daude. Zenbait autorek (Tiedemann, 2003) bestelako metodo batzuk ere bereizten dituzte (*alternative*

[1] Nolanahi ere, aitortu beharra dago corpus paraleloak ustiatzeko proposatu diren metodologietan azaldu ditugun bi helburuak ez direla beti argi bereizten. Izan ere, ingelesezko *word alignment* terminoa erabiltzen duten autore asko baliokidetza lexikalen erauzketan ari dira.

alignment models). Gehienetan, horrelakoak hibridoak izaten dira; eskuarki, lehen urratsetan elkartze-neurriak erabiltzen dituzte, eta gero estimazio-metodoren iteratibo bat erabili ohi dute emaitzak hoberentzeko.

Elkartze-neurriek (*AM, associations measures*) bi gertakariren arteko menpekotasuna neurtzen dute. Bestela esanda, AMen bidez, hitz-bikote edo 'bigrama' multzo batetik (w_1, w_2) korrelazio handia duten bikoteak identifikatzen dira. 'Korrelazio' diogunean, elkarrekin agertzeko edo gertatzeko 'joeraz' ari gara, hau da, hitz bakoitzaren maiztasuna kontuan harturik, zoriz legokiekeen baino maiztasun handiagoz agertzea elkarren ondoan (Evert, 2005).

Esaldi mailan parekatutako corpus paraleloen kasuan, bigrama hau da: hizkuntza bateko esaldi edo 'segmentu' bateko unitate batez eta beste hizkuntzako segmentuko unitate batez osatutako bikotea. Printzipioz, segmentu bakoitzeko unitate baten baliokide hautagaia beste segmentuko edozein unitate izan daiteke.

Horiez gain, zenbait heuristiko eta prozedura ere konbinatu ohi dira metodo horiekin:

- Kognatuak (edo 'sustraikideak'): antz handia duten hizkuntza desberdinetako hitzak dira kognatuak (etimologia berekoak, kultur hitzak; maileguak). Adibidez:

  biomasa / biomasa; ictioplancton / iktioplankton; arrastre / arraste

  Bi hitzen arteko kognatu-neurria kalkulatzeko, antza (*string similarity*) neurtu behar da. Horretarako, Dice koefizientea edo LCSR (*longest common sub-sequence ratio*) erabili ohi dira (Tiedemann, 2003). Bestetik, hizkuntza bakoitzaren grafiak eragindako desberdintasunen eragina gutxitzearren, arau fonologikoak erabil daitezke. Gaztelania eta euskararen artean, adibidez:

  -c- ⇔ *-k-*; *-ns-* ⇔ *-nts-*; *r-*⇔ *err-*

- Itzulpen-unitatearen bi segmentuak hautagai bakar banaz osatuak izatea: izenburuak, zerrendetako itemak…; horrelakoak baliokide 'segurutzat' jo daitezke, eta erauzte-prozesuaren lehen urratsean balioetsi

- Teknika linguistikoak: helburua izan daiteke, batetik, hizkuntza bakoitzeko hautagaiak ezaugarri morfosintaktikoaren arabera zedarritzea; bestetik, bi hizkuntzetako hautagaien arteko baliokidetzak kategoriaren edo eredu morfosintaktikoaren iragazkitik igaroaraztea
- Posizioa kontuan hartzea: euskara eta gaztelaniaren kasuan, ez du informazio baliagarria ematen, hizkuntza horien ordena kanonikoa aski desberdina baita
- Corpusetik kanpoko baliabideak erabiltzea (hiztegiak): prozesamendu-denbora gutxitzeko, baliabide horietan ez dauden baliokidetza berrien bilaketan kontzentratzeko…

Aipatu beharra dago, azkenik, hitz anitzeko unitateen arazoa. Sistema askotan, hasierako urratsa hitz bakunen arteko baliokidetzak erauztea izaten da. Hor geldituz gero, prozedura kamutsa da, hizkuntzan hain ohikoak diren hitz anitzeko unitateak aintzat hartzen ez direlako; are gehiago terminologiaz ari garela, non terminoen erdia baino gehiago hitz anitzekoak baitira (Alegria *et al*., 2004b).

Elkartze-neurrien bidezko sistemetan, bi teknika nagusi erabili dira:

- Hizkuntza bakoitzeko hitz anitzeko unitateak aldez aurretik identifikatu, eta ondoren parekatzea
- Hitz anitzeko unitateak parekatu ahala identifikatzea ('on the fly')

Beste batzuetan (Melamed, 2001), hitz bakunen arteko hiztegi probabilistiko batetik abiatuta, sistemak ondoz ondoko hitz-multzoak osatu eta probabilitatea edo elkartze-neurri jakin bat berrestimatzen du, urratsez urrats, maximizatu arte.

6 Helburua

Corpus paraleloetatik baliokidetza lexikalak erauzteko lanen azpimultzo bat landu dugu:

- Baliokidetza-mota: terminologia
- Corpus-mota (lehengaia): itzulpen-memoriak (TMX)

Termino-motako baliokideez diharduganean, honelako baliokidetza-motez ari gara:

- Hitz bakunak (*palangre* ⇔ *tretza*)

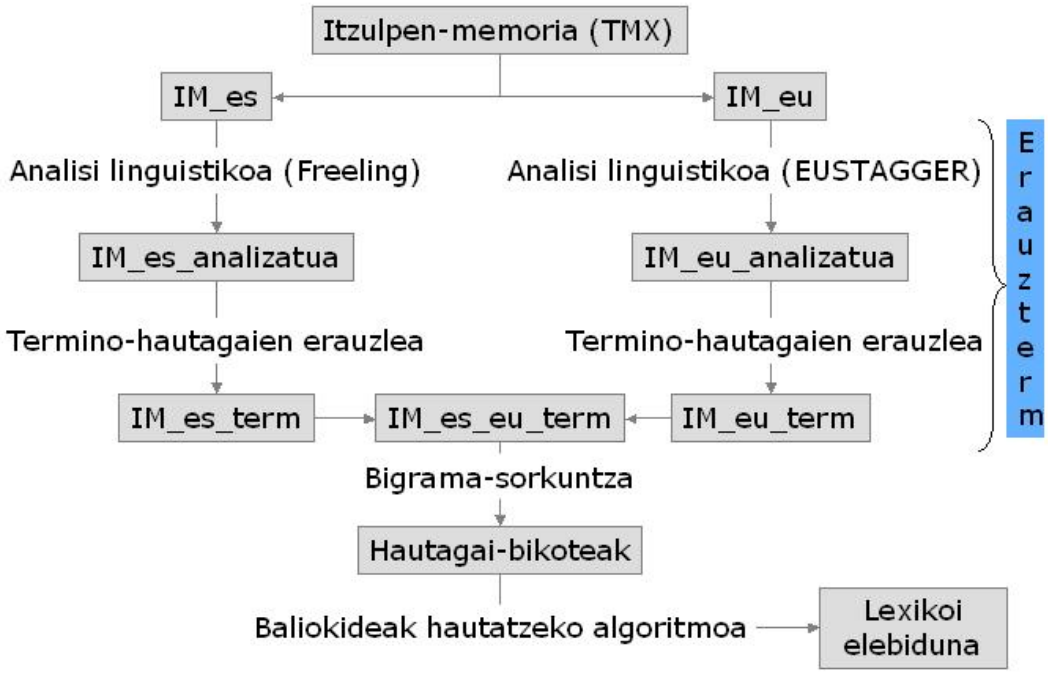
- Hitz anitzeko unitateak
    - Sintagma-unitate terminologikoak (kontzeptu-adierazleak): *arte de cerco* ⇔ *inguraketa-sare*; *rendimiento máximo sostenible* ⇔ *gehienezko errendimendu jasangarri*

7 Metodologiaren azalpen orokorra

Lehenik, erauzte-estrategia bat hautatu da. Honakoa da horren garapena:

- Hautagaien aurretiko erauzketa elebakarra, eta gero baliokideak bilatzea
    - Hizkuntza bakoitzeko hautagaiak erauztea
    - Segmentu bereko hautagaien bigramak sortzea (baliokidetza-hautagaien bikoteak)
    - Bigramen baliokidetza-neurriak kalkulatzea
        - Elkartze-neurriak (MI, Dice, t neurria, khi karratua, egiantz-arrazoia)
        - Kognatu-neurria
    - Baliokide-bikote onenak hautatzeko algoritmoa egikaritzea

Beraz, gure estrategiaren lehen ezaugarri nagusia izan da hizkuntza bakoitzeko termino hautagaiak bereiz detektatzea, eta, ondoren, itzulpen-memoriako `<tu>` unitate bereko segmentuetan ageri direnak konbinatzea, hautagai-bikoteak lortzeko; gero, hautagai-bikote bakoitzaren informazioa (elkartze-neurriak eta kognatu-neurria) konputatu, eta, azkenik, informazio hori erabiliz, baliokide-hautagai onenak hautatzeko algoritmoa aplikatzen da. Ondoren, erauzte-estrategia hori gauzatzeko urratsez urratseko prozesua zehaztu da. Hau da erauzte-prozesuaren diagrama orokorra:



4. irudia. Baliokidetzak erauzteko prozesua.

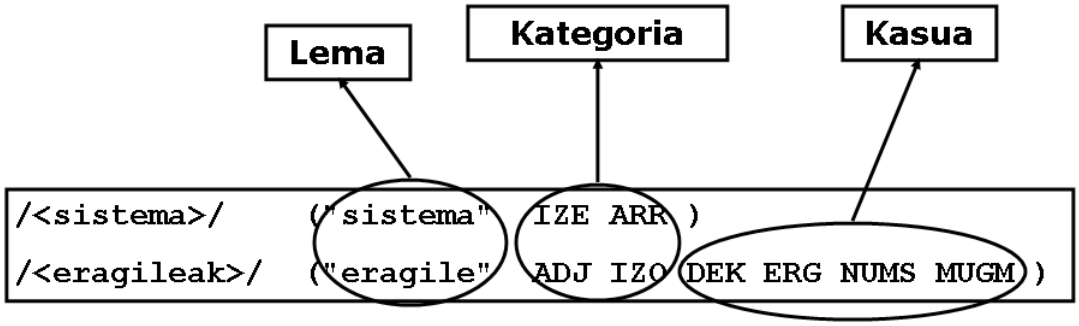
Hurrengo ataletan, prozesu horren urratsak xeheago azalduko ditugu.

8 Erauzte-prozesua

Lehen urratsa da itzulpen-memoriaren hizkuntza bakoitzeko atalak bereiz prozesatzeko bereizketa. Hori egiteko, XPath lengoaia erabiltzen duten Perl scriptak egin dira. Hortik aurrera, atal bakoitzaren prozesamenduari ekiten zaio.

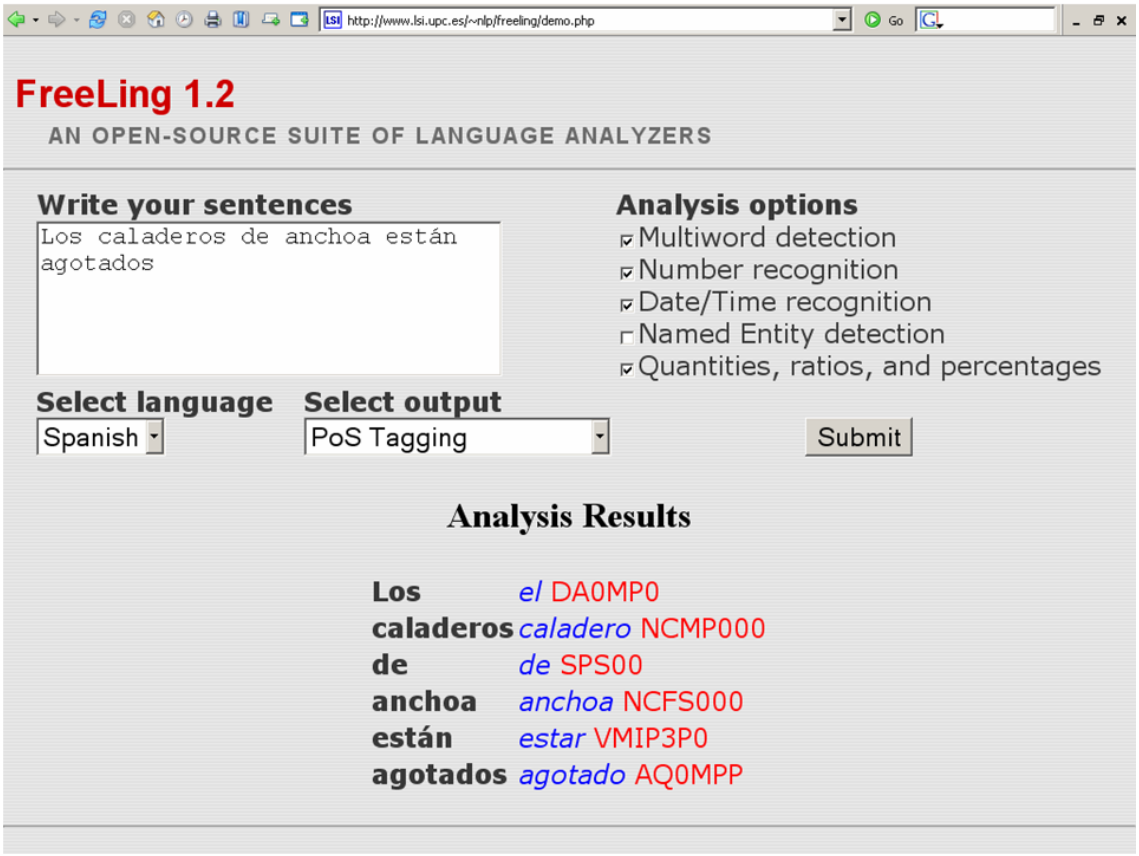
8.1 Itzulpen-memorien prozesatze linguistikoa

Euskarazko testua prozesatzeko, IXA taldeak garatutako EUSTAGGER lematizatzaile/etiketatzailea erabili dugu. EUSTAGGERek eskaintzen duen informazio linguistikotik, lema, kategoria eta kasua gorde dira. Informazio hori beharrezkoa da hurrengo urratsean, termino hautagaien erauzketan, eredu morfosintaktikoaren araberako hitz-segidak (termino hautagaiak) detektatzeko eta hautagai horien forma kanonikoak (*forma* lema* erako sintagmak) emateko.



5. irudia. EUSTAGGERen informazioa.

Gaztelaniazko testua prozesatzeko, UPCko *Centre de Tecnologies i Aplicacions del Llenguatge i la Parla* (TALP) eta Bartzelonako Unibertsitateko *Centre de Llenguatge i Computació* erakundeek garatutako *Freeling* software libreko paketea erabili dugu (Carreras *et al.*, 2004) (http://garraf.epsevg.upc.es/freeling/). *Freeling*-ek testu-hitzen (*token*-en) lema eta kategoria ematen ditu lehen urratsean (analisi morfologikoan), eta, analisi bat baino gehiago dagoenean, analisi bakoitzaren probabilitatea ere ematen du. Beheko pantailan, 'POS tagging' (kategoria-etiketatzea) aukera erabili da lema eta kategoria bikote bakarra esleitzeko testu-hitz bakoitzari.

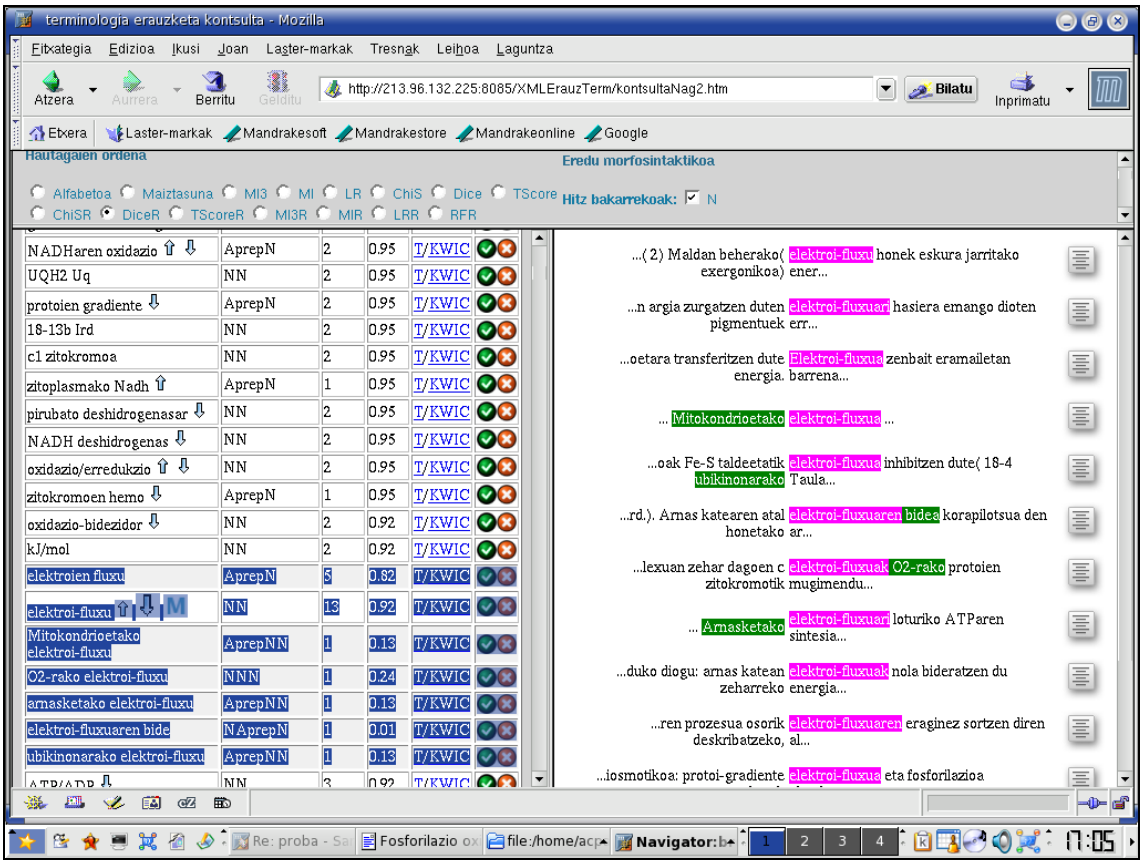


6. irudia. *Freeling* etiketatzailearen informazioa.

8.2 Termino hautagaien detekzioa

8.2.1 Euskarazko hautagaiak: *Erauzterm*

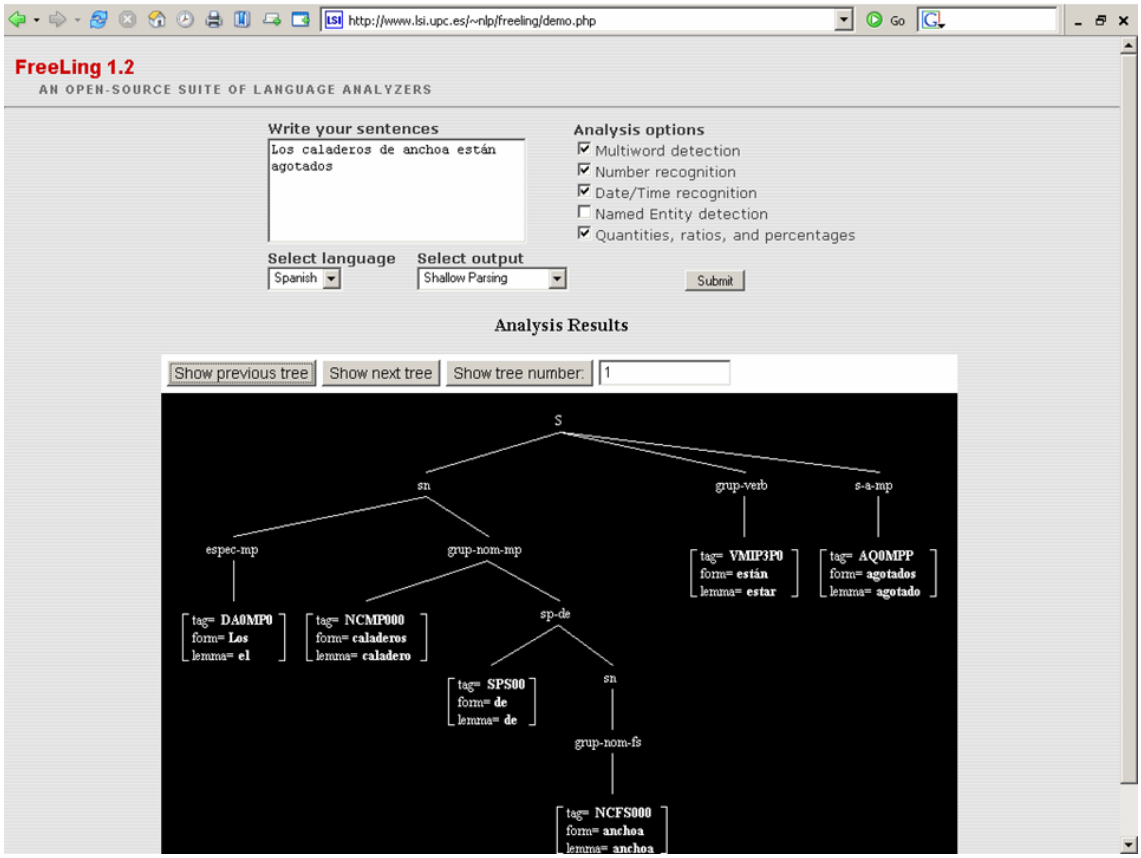
Euskarazko termino hautagaiak detektatzeko, HIZKING21en garatu dugun *Erauzterm* prototipoa erabili dugu (Alegria *et al.*, 2004a, 2004b, 2005). *Erauzterm*-ek analisi linguistikoa eta estatistikoa hartzen ditu bere baitan, eta termino hautagaiak, horien forma kanonikoa, eredu morfosintaktikoa eta informazio estatistikoa ematen ditu. Horren bidez, itzulpen-memoriaren euskarazko segmentu bakoitzean egon daitezkeen termino-hautagaiak erauz ditzakegu.



7. irudia. *Erauzterm*-en interfazea.

8.2.2 Gaztelaniazko hautagaiak: *Freeling* (<grup-nom>)

'Shallow parsing' (azaleko analisia) aukera erabiliz, esaldiaren zuhaitz-egitura ematen du *Freeling*-ek. Horrek aukera eman digu izen-sintagmak markatzeko, <grup-nom> etiketa duten sintagmak hartuz. Beheko adibidean, *caladero de anchoa* izen-sintagma markatuko litzateke.



8. irudia. *Freeling*-en analisi sintaktikoaren emaitzak.

8.2.3 Termino habiatuen tratamendua

Terminoak IS luzeago baten barnean 'habiaturik' egon daitezke: *datu-base lexikal* / *datu-base merke; reclutamientos recientes* / *reclutamientos medios; esfuerzo efectivo* / *esfuerzo dirigido.*

Termino habiatuak ere hautagai izateko, prozedura hauek inplementatu ditugu.

- *Erauzterm*: IS luzeenak deskonposatzea
  - Eredu morfosintaktikoak azpisintagmatan banatzea
  - Eredu batekin bat datozen burua edota modifikatzailea hautatzea
- *Freeling*: <grup-nom> habiatuak

8.3 Hautagai-bikoteen sorkuntza

Linguistikoki prozesatutako hizkuntza bakoitzeko memoria-atalak batu ondoren, `<tu>` itzulpen-unitate bereko hizkuntza bakoitzeko hautagaiak konbinatu egiten dira, hautagai-bikoteak edo bigramak osatzeko. Horiek datu-base erlazional batean

biltegiratzen dira, eta ondoren bikote bakoitzaren informazioa kalkulatzen da: elkartze-neurriak eta kognatu-neurria.

Hitz anitzeko hautagaien bikoteen kognatu-neurria kalkulatzeko, karaktere-kate osoak kontuan hartu ohi dira, eta gerta daiteke, hizkuntzen izaerak hartaratuta, osagaien ordena bera ez izatea. Adibidez, *Inteneteko konexio / conexión a Internet*. Egokiagoa dirudi, hortaz, osagai bakunen arteko kognatu-neurria kalkulatzea, eta kognatu-neurri handieneko konbinazioa hautatzea termino osoen artekoa kalkulatzeko, behar diren ordena-aldaketak eginda.

8.4 Baliokide onenak hautatzeko algoritmoa

Hautagai-bikote bakoitzaren informazioa bildutakoan, algoritmoaren zeregina da, informazio hori erabiliz, hautagai-bikote onenak hautatzea. Baliokidetza lexikalak erauzteko egin diren lan gehienetan, oinarrizko hipotesia izan da corpus paraleloan hizkuntza bateko hitzek itzulpen edo baliokide bakarra dutela beste hizkuntzan (Fung, 1998). Autore gehienek aitortzen dute hipotesi hori oso gutxitan betetzen dela, baina hurbilketa egokia izan daitekeela helburu askotarako, batik bat esparru mugatu edo espezializatuetan erabili nahi diren sistemetarako. Horiek horrela, eta hurbilketa mugatua dela jakinik ere, jatorrizko hizkuntzako hautagai batek hautagai-bikoteen datu-basean dituen baliokide posible guztietatik 'onena' hautatzea izan da erabili den algoritmoaren estrategia.

Oro har, bi algoritmo-mota erabili dira (Matsumoto *et al.*, 2000):

- EM algoritmoan oinarrituak (*Expectation Maximization*): iteratiboak
- Urratsez urratseko algoritmoak ('greedy' algoritmoak): urrats bakoitzean 'hautatzen' diren baliokideak ez dira sartzen hurrengo urratsean; batzuetan, neurriak berriz kalkulatzen dira

Lehen motako algoritmoak itzulpen automatikoko sistema estatistikoetan erabili izan dira batez ere (Brown *et al*. 1993; Dagan *et al.*, 1993). Ildo horretako garapen ezagunenetakoa GIZA++ sistema da (Och *et al.*, 2000).[2] Elkartze-neurrien bidezko sistemetan ere saiakuntzak egin dira (Kupiec, 1993; Hiemstra, 1999, Melamed, 2001).

Proiektu honetan, urratsez urratseko algoritmo bat erabili dugu. Horrelakoen

[2] GIZA++ eu-es memoriak esaldia baino txikiagoko unitateetan parekatzeko nola aplikatu den jakiteko, ikus Nevado *et al.*, 2004.

abantaila nagusia sinpletasuna eta azkartasuna da; desabantaila izan daiteke, ordea, urrats batean hautatzen diren baliokidetzak ez direla berraztertzen. Hau da inplementatu den algoritmoaren eskema:

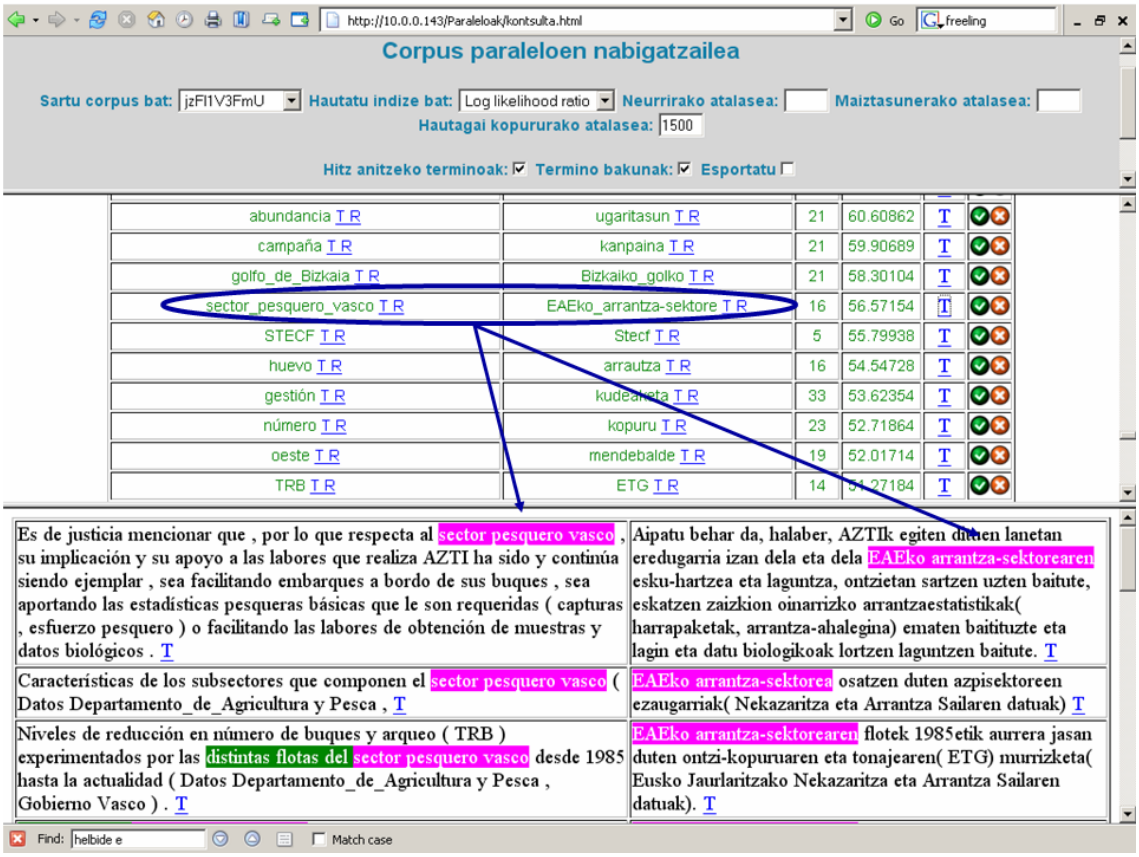
- 1:1 segmentuak (baliokide hautagaiek segmentu oso bana hartzen dutenean)
- Kognatu-neurri handiko bikoteak (LCSR > 0,8)
- Gainerako bikoteak: elkartze-neurrien balioen araberako *competitive linking algorithm* (Melamed, 2001) edo *meilleur affectation biunivoque* (Kraif, 2002a) prozedurak:
  - Balio handieneko elkartze-neurria duen es_i-eu_j bikotea hartu eta egiaztatu eu_j-es_i ere balio handienekoa dela
  - es_i eta eu_j horien gainerako bigramak baztertu
  - Hurrengo balio handieneko bikotea hautatu eta aurreko urratsak egin, bigramen-multzoa hustu arte

9 Erabiltzailearen lan-interfazea

Erabiltzailearen interfazearen helburua da erabiltzaileak erauzketaren emaitzak aztertzeko eta kudeatzeko aukera izatea. Erabiltzaileari eskaintzen zaizkion aukerak:

- Erauzketa-corpusa hautatzea
- Emaitzak hainbat neurri estatistikoren arabera bistaratzea
- Zenbait atalaseren arabera iragazteko aukera: neurri estatistikoa, maiztasuna, baliokide-kopurua
- Baliokidetzen testuingurua bistaratzea
- Baliokidetzak balioesteko eta esportatzeko aukera

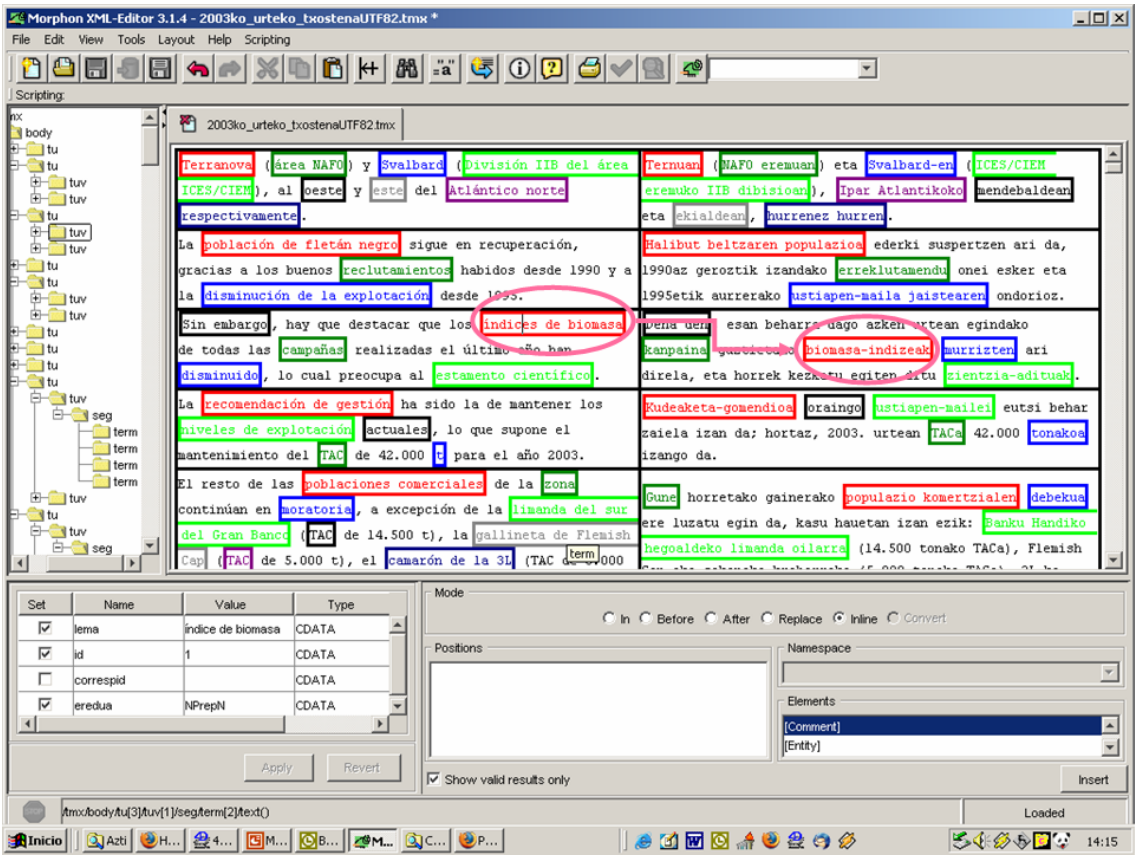
Honelakoa da diseinatu den erabiltzaile-interfazea:



9. irudia. Erauzketaren emaitzak ikusteko eta lantzeko interfazea.

10 Ebaluazioa

Saiakuntzak bi corpusekin egin ditugu: AZTI (1.889 segmentu; es: 36.990 hitz; eu: 25.589 hitz); Euskaltel (10.900 seg.; es: 153.163 hitz; eu: 110.165 hitz). Corpus horietatik automatikoki egindako erauzketa ebaluatzeko, baliokidetzak eskuz landu dira. Erauzketa XML editore batean egin da (Morphon XML-Editor 1.4). Hau da prestatu dugun erauzte-interfazea:



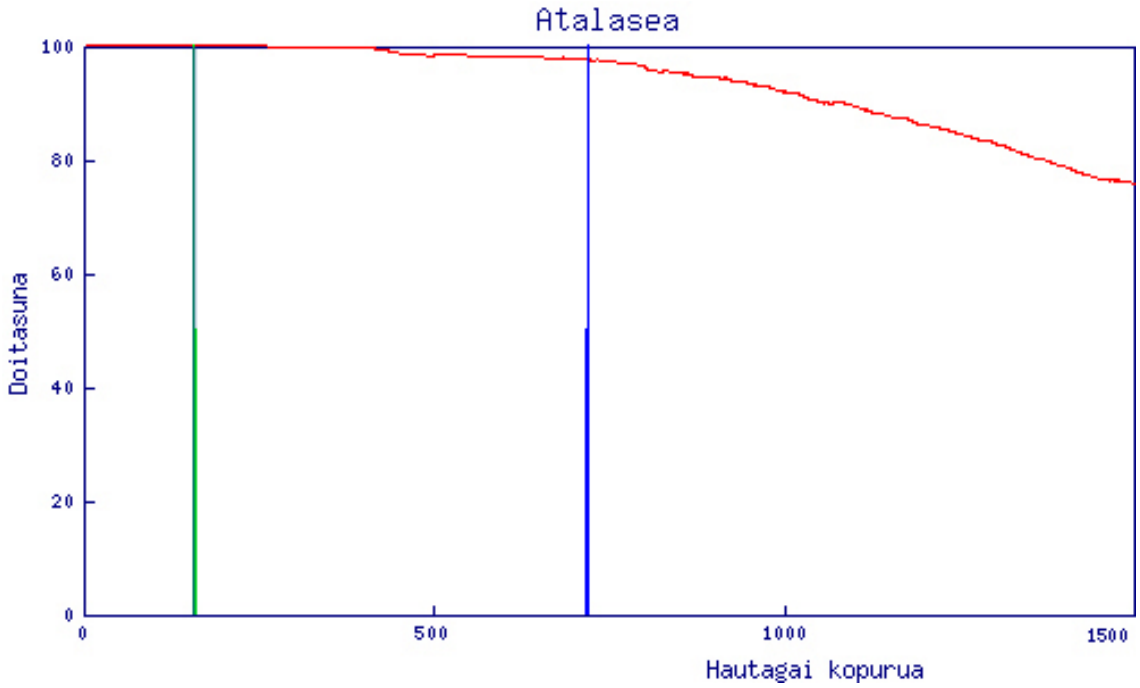
10. irudia. IMetatik baliokidetzak eskuz markatzeko lan-interfazea.

Esku-erauzketa egitean, kontuan hartu behar dira 8.4 atalean onartu dugun baliokide bakarraren lan-hipotesiak ezartzen dituen mugak. Esaterako, AZTIren itzulpen-memorian gaztelaniazko *pesquería* terminoa bi kontzepturen adierazlea da eta hauek dira euskarazko baliokideak: jarduera = *arrantza*; eta tokia = *arrantza-toki, arrantza-leku*. Baliokide bakarraren hipotesiak ezin ditu polisemia- eta sinonimia-erlazioak tratatu, ezta termino-aldaerak ere; beraz, ebaluazioak ezarritako helburuei erantzun diezaion, eskuz erauzten diren baliokidetzetatik bakarra hartu da kontuan: maiztasun handienekoa. Tresnaren etorkizuneko garapenean, muga hau gainditzeko prozedura ikertu eta inplementatuko dugu. Esku-erauzketan agerian gelditzen dira, halaber, 5.1 atalean azaldu ditugun arazoak: itzultze-baliokidetzak termino-mailakoak ez izatea (*En 2001 los desembarcos superaron las 2.200 t. ⇔ 2001. urtean 2.200 tona baino gehiago lehorreratu ziren*), elipsiak, itzuli gabeko pasarteak...

11 Emaitzak

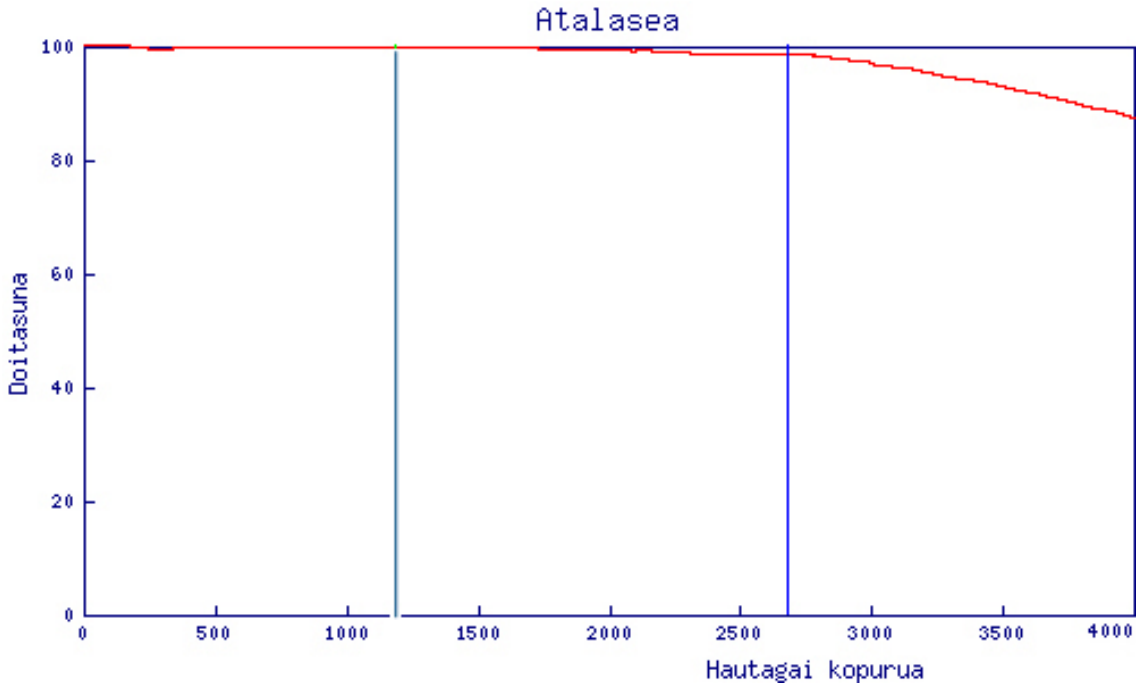
Bi corpusak automatikoki prozesatu ondoren, erauzketaren emaitza aztertu dugu,

eta zuzentzat jo ditugun baliokide-bikoteak balioetsi. Hurrengo bi irudietan ikus daiteke nola aldatzen den doitasuna (bikote zuzenen ehunekoa) tresnak ematen duen baliokidetza-kopuruaren arabera. Doitasuna neurtzean, *Freeling*-ek ematen duen lema-egitura hartu da kontuan, eta ez terminoaren berezko forma kanonikoa.[3] Ebaluazioan erabili diren elkartze-neurri estatistikoetatik, egiantz-arrazoiak (LR) izan ditu emaitza onenak, eta horien araberako emaitzak irudikatu ditugu. Lerro bertikalek algoritmoaren urratsen arteko mugak adierazten dituzte; hurrenez hurren, lehenak 1:1 segmentuen urratsean lortzen diren balokideen amaiera, eta bigarrenak kognatu-neurrien bidezko erauzketak itzultzen dituzten baliokideen amaiera; hortik aurrera, elkartze-neurrien bidezko baliokideak.



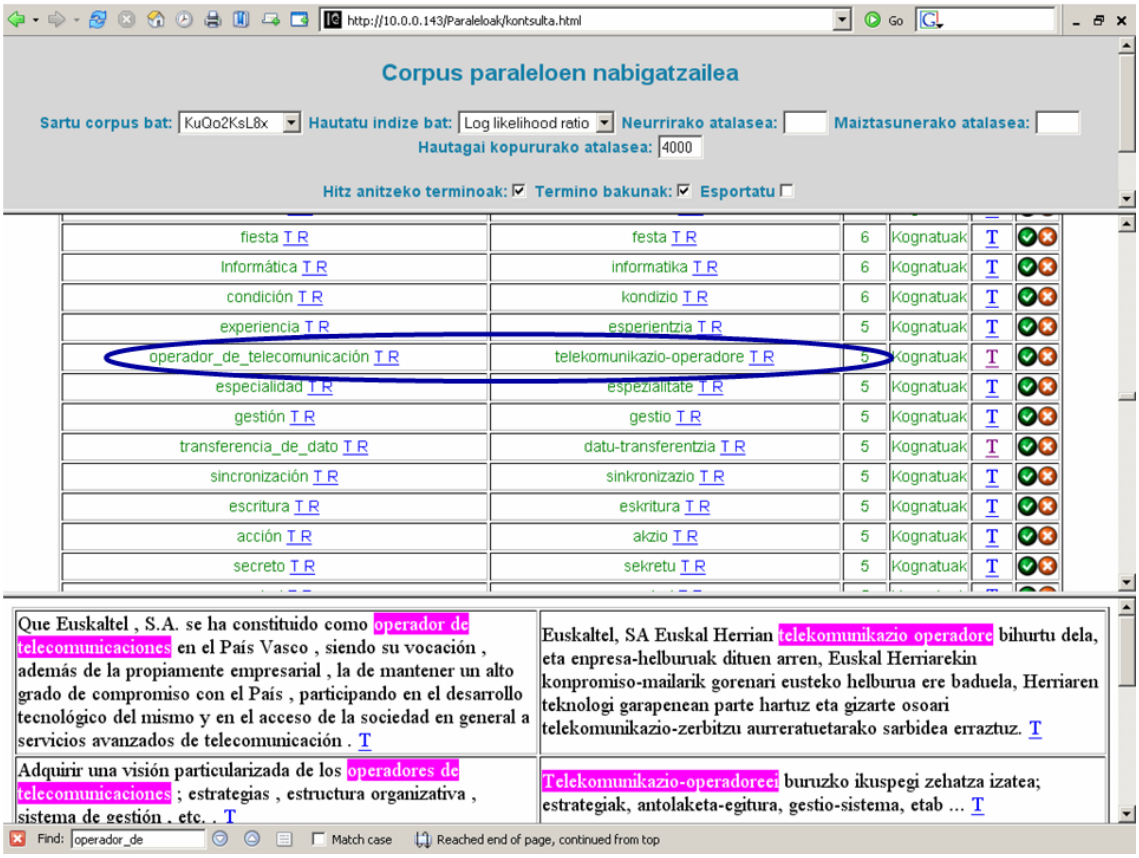
11. irudia. AZTIren itzulpen-memoriaren erauzketa automatikoaren emaitzak.

[3] Adibidez, *bases de datos* testu-formaren terminoaren forma kanonikoa *base de datos* da. Hau da, lehen osagaiaren flexioak ez du eraginik forma kanonikoan, baina bigarrenarenak bai. Hala ere, *Freeling*-etik lortzen den sintagma lematizatua *base de dato* da. Proiektuaren hurrengo garapenetan, aurre egingo diogu forma kanonikoa doiago erauzteari; horretarako, terminoaren ereduaren araberako erregelak formulatuko dira, eta corpuseko bertako informazio estatistikoa ere erabiltzeko asmoa dago. Euskararen kasuan, ez dago horrelako arazorik, flexioa beti amaieran dagoelako eta terminoak *forma* lema* moduan erauzten direlako zuzenean.



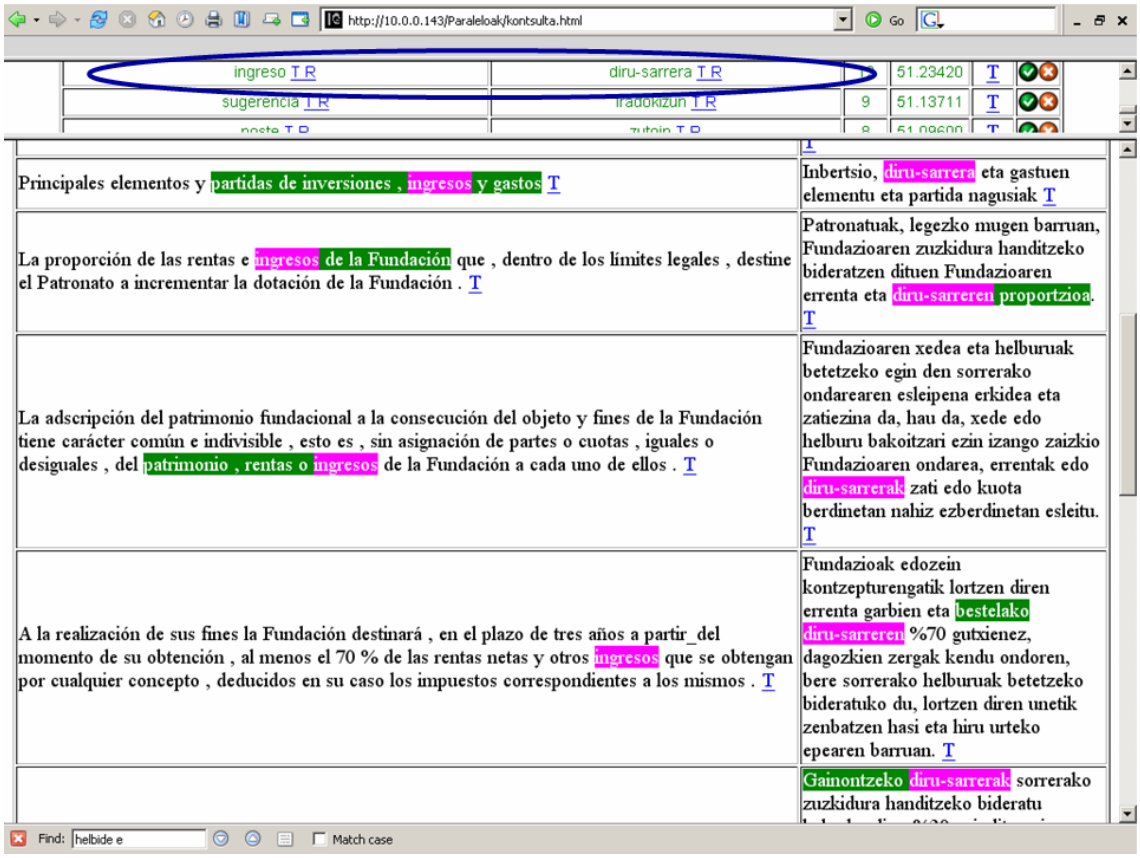
12. irudia. Euskaltel-en itzulpen-memoriaren erauzketa automatikoaren emaitzak

Doitasun-datuak aski onak dira. 1:1 segmentuetako hautagaiak, aurrez ikusi dugun bezala, oso baliokide seguruak dira; bestetik, bistan da emaitza onak lortzen direla kognatu-neurria > 0,8 duten hautagaiak baliokidetzat jota. Hurrengo adibidean, argi ikusten da hitz anitzeko terminoen kognatu-neurria kalkulatzeko proposatu dugun sistema egokia dela alderantzizko ordenako terminoetarako ere:



13. irudia. Kognatuen erauzketa.

Bestetik, erauzte-estrategia egokia da osagai-kopuru desberdineko termino baliokideak erauzteko. Hurrengo adibidean, 1:2 luzera-erlazioko adibide bat dago ikusgai (*ingreso* ⇔ *diru-sarrera*):



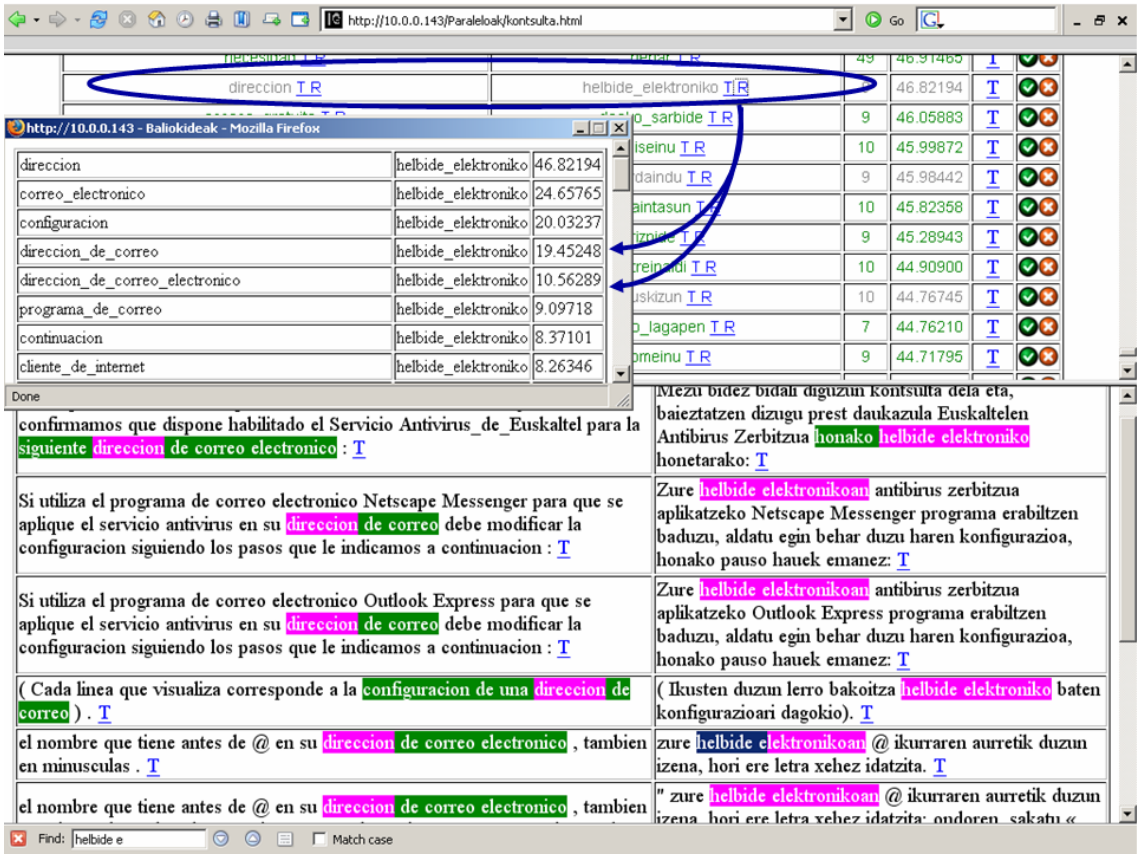
14. irudia. n:m baliokidetzen erauzketa.

Estaldura-datuak, berriz, apalagoak dira (% 50 ingurukoak).[4] Horren arrazoi posibleak:

- Estalduran eragina duten prozesu-kate bat dago:
  - Hizkuntza bakoitzeko analisi linguistikoan egiten diren lematizazio edo kategoria-esleitze desegokiak
  - Hizkuntza bakoitzeko hautagaien detekzioan atzematen ez diren terminoak; adibidez, euskarazko termino-eredu konplexu batzuk ez daude gramatikan; gaztelaniaz, Freeling-ek *de* ez diren preposizio-sintagmak ez ditu aurreko <grup_non> sintagmen azpian sartzen; horretara, *mortalidad por pesca* ez du termino-hautagaitzat jotzen, *mortalidad* hutsa baizik; horren ondorioz, *mortalidad* ⇔ *arrantza-hilkortasun* moduko baliokidetzak erauzten dira. Horrek eragina du estalduran zein doitasunean

[4] Nolanahi ere, argitu behar da estaldura hori doitasun zehatz baterako dela, alegia, estaldura handitu daitekeela, doitasunaren kaltetan.

- Esku-erauzketaren arazoez mintzatu garenean azaldu ditugun arazo batzuek eragina dute estalduran: IS ez diren termino-ereduak, juntadura, elipsia...
- Hizkuntza bateko habiatuak elkartze-neurri handiagoa eman dezake hautagai osoak baino (hizkuntza batean termino bakarra izaki, bestean bat baino gehiago daudenean, baina osagai bat komuna denean). Adibideak: *cuota mensual* ⇔ *hileko/hileroko/hilabeteko kuota*; erauzketan hautagai osoak (sintagma luzeenak) eta habiatuak kontuan hartzen direnez, elkartze-neurri handiena osagai komunak eman dezake (*kuota*-k), eta baliokidetza 'onena' hau izan daiteke: *cuota mensual* ⇔ *kuota.* Hurrengo irudian, arazo horren adibide bat dago. Erauzi den baliokidetza *dirección* ⇔ *helbide elektroniko* da. Euskarazko *helbide elektroniko* terminoaren eskuinean dagoen 'R' estekan sakatuta, horren baliokide hautagai guztiak bistaratzen dira. Bistan da gaztelaniaz *dirección de correo* eta *dirección de correo electrónico* direla testuko baliokideak. Habiatuak kontuan hartu ditugunez, *dirección* da elkartze-neurri handiena ematen duena:



15. irudia. Euskarazko *helbide elektroniko* terminoaren baliokide hautagaien

informazioa.

- Zeharkako elkartzearen fenomenoa (*indirect association*; Melamed, 2001). Arazo hau hitz bakunen erauzketarekin lotu izan da, eta kolokazioen eragina izaten da. Kolokazio-fenomenoa izan gabe ere, gerta daiteke, batez ere corpusa oso handia eta heterogeneoa ez bada, bi hitz segmentu gehienetan elkarrekin agertzea, eta berdin gertatzea beste hizkuntzako baliokideekin. Horien arteko baliokidetza gurutzatuak gerta daitezke

Gure ondorioa da estaldura handitzeko lehen neurriak prozesu linguistikoetan hartu behar direla (euskarazko eredu morfosintaktikoen multzoa handitzea; *Freeling*-en irteera prozesatzea; aldaeren tratamendua..).

12 Ondorioak

Itzulpen-memorietatik terminologia elebiduna automatikoki erauzteko prototipoa garatu dugu. Prototipoak prozesu osoa automatizatzen du: itzulpen-memoria TMX formatuan kargatzen denetik, erabiltzailearen interfazean emaitzak bistaratu arte. Sistemak automatikoki egiten ditu hasierako eta amaierako urrats horien arteko urrats guztiak.

Erabiltzailearen interfaze erabilgarri eta intuitibo bat ere inplementatu da. Horri esker, erabiltzaileak erauzketaren emaitzak azter ditzake, eta aukera du hainbat neurri estatistikoren araberako emaitzak lortzeko eta zenbait atalaseren arabera iragazteko (neurri estatistikoa, maiztasuna, baliokide-kopurua); horrekin batera, baliokidetzen testuingurua bistaraz dezake, eta baliokidetzak balioesteko eta esportatzeko aukera ere badago.

Erauzte-prozesuaren emaitzen ebaluazioak erakutsi du, batetik, estaldura hobetu beharra dagoela, eta, bestetik, urratsez urratseko algoritmoaren portaera ere hobetzea komeni dela. Hobekuntza horiek nola egin daitezkeen ere aurreikusi da, eta proiektuaren etorkizuneko garapenean ekingo zaie lan horiei.

Bibliografia

Abaitua, J. 1997. Segmentos y unidades de traducción. . [http://www.serv-

inf.deusto.es/abaitua/konzeptu/l2uutts.htm; 06-02-14an irakurria]

Alegria, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S. & Urizar, R. 2004a. "Linguistic and Statistical Approaches to Basque Term Extraction." In *GLAT-2004: The Production Of Specialized Texts.* [http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1079630425/publikoak/Term_Erauzketa.pdf; 06-02-14an irakurria]

Alegria, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S. & Urizar, R. 2004b. "A Xml-Based Term Extraction Tool for Basque." In *LREC2004: 4 Th International Conference On Language Resources And Evaluation.* [http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1078851980/publikoak/Erauzterm_LREC; 06-02-14an irakurria]

Alegria, I., Gurrutxaga, A., Saralegi, X., & Ugartetxea, S. 2005. "Erauzterm: euskarazko terminoak erauzteko tresna erdiautomatikoa." In *Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera zientifiko-teknikoa*. [http://www.mendebalde.com/modulos/usuariosFtp/conexion/archi146A.pdf; 06-02-14an irakurria]

Brown, P.F., Della Pietra, S., Della Pietra, V.J. & Mercer, R.J. 1993. "The Mathematic of Statistical Machine Translation: Parameter Estimation." In *Computational Linguistics* 19-2. 263-311. orr.

Carreras, X., Chao, I., Padró, L., & Padró, M. 2004. "FreeLing: An Open-Source Suite of Language Analyzers" In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisboa.

Casillas, A., Fernández, I. & Martínez, R. 2004. "Sentence alignment for spanish-basque bitexts: word correspondences vs. markup similarity." In *Fifth International Conference on Intelligent Text Processing and Computational Linguistics – CICLing*. Seul.

Dagan, I., Church, K. & Gale, W. 1993. "Robust bilingual word alignment for machine aided translation." In *Proceedings of the Workshop on Very Large Corpora (WVLC)*. 1-8. orr.

Estopá, R. 1999. *Extracció de terminologia: elements per a la construcció d'un*

SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada). Doktore-tesia. Bartzelona: IULA-Universidad Pompeu Fabra.

Evert, S. 2005. *Computational Approaches to Collocations*. [www.collocations.de; 06-02-14an irakurria]

Fung, P. 1998. "A Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora". In *Lecture Notes in Artificial Intelligence AMTA 98.*, Springer Publisher, 1998, vol 1529, 1-17. orr.

Hiemstra, D. 1996. *Using statistical methods to create a bilingual dictionary.* Master's Thesis, University of Twente. [http://wwwhome.cs.utwente.nl/~hiemstra/papers/hiemstra96.pdf; 06-02-14an irakurria]

Kraif, O. 2002a. "Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné." *Alignement lexical dans les corpus multilingues, Lexicometrica,* Jean Véronis ed., [on line] [05-06-10] [http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm; 06-02-14an irakurria]

Kraif, O. 2002b. "Translation aligment and lexical corrspondence." In: Altenberg, B. & Granger, S. (ed.) *Lexis in Contrast.* Amsterdam: John Benjamins.

Kupiec, J. 1993: "An Algorithm for finding Noun Phrase Correspondences in Bilingual Corpora." In *31st Annual Meeting of the Association for Computational Linguistics* 17-22. Columbus, Ohio.

Matsumoto, Y. & Utsuro, T. 2000. "Lexical Knowledge Acquisirtion." In DALE, R. et al. (ed.). *Handbook of Natural Language Processing*. New York/Basel: Marcelo Dekker, Inc. 563-610 orr.

Melamed, I.D. 2001. *Empirical Methods for Exploiting Parallel Texts.* MIT Press.

Merkel, M & Ahrenberg, L. 1999. *Evaluating Word Alignment Systems*. PLUG report. [http://www.ida.liu.se/~magme/publications/eval-plug.pdf; 06-02-14an irakurria]

Nevado, F., Casacuberta, F. & Landa, L. 2004. "Translation memories enrichment by statistical bilingual segmentation." In *Proceedings of the IV International Conference on Language Resources and Evaluation - LREC2004,* 1. orr. Lisboa.

[on line] [06-02-14] [http://prhlt.iti.es/papers/2004/Nevado04a.pdf; 06-02-14an irakurria]

Somers, H. *Bilingual Parallel Corpora and Language Engineering*. [http://www.emille.lancs.ac.uk/lesal/somers.pdf; 06-02-14an irakurria]

Tiedemann, J. 2003. *Recycling Translations. Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Doktore-tesia. Uppsala: Acta Universitatis Upsaliensis. [on line] [05-06-10] [http://stp.ling.uu.se/~joerg/phd/html/; 06-02-14an irakurria]

Vinay, J.P. & Darbelnet, J. 1958 *Stylistique comparee du francais et de l'anglaais: Methode de traduction*. Paris: Didier.