Erauzterm: euskarazko terminoak erauzteko tresna erdiautomatikoa

Elhuyar Fundazioa
A. Gurrutxaga, X. Saralegi, S. Ugartetxea

Ixa taldea–Euskal Herriko Unibertsitatea
I. Alegria

Artikulu honetan, terminologia-erauzketa automatikoaren alorrean egin ditugun lanak eta horietan oinarrituta garatu dugun tresna (Erauzterm) azalduko ditugu. Terminoak automatikoki erauzteko prozedura ugari erabili izan dira, bi teknika nagusitan oinarrituak denak ere: teknika linguistikoak eta teknika estatistikoak. Euskara hizkuntza eranskaria denez gero, metodo estatistiko hutsen emaitzak ez lirateke onargarriak. Horrenbestez, metodo hibrido baten alde egin dugu. Lehen urratsean, teknika linguistikoak erabiltzen dira termino-hautagaiak erauzteko. Bigarren urratsean, termino-hautagaiak teknika estatistikoen bidez sailkatzen dira. Garatu edota aplikatu ditugun teknika linguistikoak eta estatistikoak azaldu ondoren, prozesu osoa integratzen duen aplikazio informatikoa deskribatuko dugu. Azkenik, erauzketaren emaitzen lehen ebaluazioa egingo dugu, hau da, tresnak doitasunaren

zein estalduraren aldetik duen eraginkortasuna neurtuko da. Tresna erdiautomatikoa da, modu autonomoan lan egin baitezake, baina emaitza zehatz-zehatzak lortzeko eskuz berrikustea eta balioestea eskatzen du. Tresna horren lehen bertsioa burutu dugu, eta hobekuntzak egiten ari gara.

1. Termino-erauzketa: metodoak

Hizkuntz ingeniaritzaren beste hainbat arlotan bezala, testuetatik terminoak erauzteko ere bi bide nagusi daude: linguistikoa eta estatistikoa. Gaur egun, biak konbinatzeko joera nagusitu da, bataren zein bestearen abantailez baliatzeko asmoz. Banan-banan azalduko ditugu, haien arteko aldea eta osagarritasuna hobeto ulertarazteko.

Teknika linguistikoek terminoen egitura morfosintaktikoak edo ereduak hartzen dituzte oinarritzat (Bourigault, 1996). Egitura horien multzoa mugatua dela jotzen da, eta zehazteko modukoa ere bai. Beraz, eredu morfosintaktiko horiekin bat datozen hitzak edo hitz-segidak (sintagmak) aurkitzea eta erauztea da oinarrizko estrategia. Honelako tekniken abantaila nagusia da hasieratik beretik 'termino' izaera agertzen duten hizkuntz formetara bideratuta daudela, eta maiztasun txikikoak izanda ere, ereduen araberako termino-hautagaiak erauz daitezkeela. Lehen muga begi-bistakoa da, ordea: eredumultzo mugatu horretan benetako terminoei dagokien ereduren bat landu ez bada, aldez aurretik badakigu ezingo dela erauzi.

Teknika estatistikoetan, berriz, hitz anitzeko eta hitz bakarreko terminoak bereizi ohi dira. Hitz anitzekoen kasuan, terminoen osagaiek elkarrekin agertzeko duten 'joeraren' neurketan (probabilitateetan) oinarritzen dira (Chuck & Hanks, 1990; Smadja, 1993), aurreikusten baita testu teknikoetan terminoek maiztasun nabarmenaz agertzeko joera dutela. Hori dela eta, estatistikaren munduan ezagunak diren elkartze-neurriak (*association measures*) erabili ohi dira

horretarako. Hitz bakarreko terminoekin, berriz, bestelako neurriak erabili behar dira. Eskuarki, informazioaren berreskurapenean (*Information Retrieval*) erabiltzen diren indexazio-tekniketan oinarritutako neurriak erabili ohi dira; adibidez, termino bakunak aztergai den testuan duen maiztasun erlatiboa eta hizkuntza orokorreko datuen arabera duenaren arteko konparazioa eginez.

Batean zein bestean, metodo estatistikoen abantaila da hizkuntzatik independenteak diren sistemak garatzeko aukera ematen dutela. Horrez gain, teknika linguistikoetan behar diren baliabide lexikalak eta prozesamendu linguistikoa ez dira beharrezkoak. Desabantaila da, berriz, elkarren ondoan maiz agertuta ere termino-izaera ez duten hitz-segidak erauztea, edo, hitz bakarrekoen kasuan, berez termino ez diren hitzak, baina hizkuntza orokorreko datuekiko nabarmenak direnak, erauzten direla; hau da, emaitzak doitasun txikikoak izateko arriskua (zarata). Gainera, ezkutaturik gera daitezke termino-egitura argia duten baina maiztasun nabarmenaz agertzen ez diren hautagaiak (teknika linguistikoen bidez atzeman litezkeenak). Euskararen kasuan, gainera, testu-hitzen agerraldien datuak sakabanatuagoak dira lemen edo terminoen forma kanonikoenak baino, eta aurrez esan daiteke emaitzak ez direla gogobetegarriak izango.

Sistema gehienetan, linguistikoetan nahiz estatistikoetan, hitz anitzeko terminoak bilatu ohi dira, hitz bakarrekoak polisemikoagoak dira eta (Bourigault & Jacquemin, 1999):

"On the one hand, term extractors focus on multi-word terms for ontological motivations: single-word terms are too polysemous and too generic and it is therefore necessary to provide the user with multi-word terms that represent finer concepts in a domain."

Horrek ekarri du berekin hitz anitzeko terminoen erauzketaz diharduten lanak askoz ugariagoak izatea literaturan, hitz bakarrekoen erauzketari buruzkoak baino.

Azken urteotan, metodo hibrido edo konbinatuak nagusitu dira, eta teknika linguistikoak zein estatistikoak batera erabiltzen dira, elkarren osagarri. Konbinazio hori bitara egin daiteke: a) lehenik teknika estatistikoak erabiltzea, eta, termino-hautagai esanguratsuenak lortu ondoren, eredu morfosintaktikoen bidez hautatzea (Samdja, 1993); b) teknika linguistikoen bidez hautagai-multzoa zehaztea aurrena, eta gero teknika estatistikoak aplikatzea, hautagaiak ordenatzeko edota iragazteko (Daille, 1995; Justeson, 1993).

Gure proiektuan, azken prozedura horren alde egin dugu, hizkuntzaren ezaugarriek hartaraturik batez ere; beraz, euskarazko terminoen ezaugarriak identifikatzea izan da gure lehen eginkizuna.

Terminoen erauzketa erabat automatikoa izan daiteke –informatikaren arlo klasikoa den informazioaren berreskurapenean (*Information Retrieval*) erabiltzen da batzuetan– baina, aplikazio gehienetan, modu erdiautomatikoan lan egiten da: programak proposamen bat egiten du eta erabiltzaileak gainbegiratzen du emaitza. Horrelako tresna bat onuragarria izan daiteke euskarazko terminologia aztertzeko eta finkatzeko, oso laguntza bizkor eta zehatza eskain baitezake zein termino erabili den eta zein maiztasunez jakiteko.

Azken hamabost urteetan euskararen prozesamendu automatikoan egindako aurrerapenen ondorioz, gaur egun aukera dugu euskararen analisi morfologikoa, lematizazioa eta azaleko analisi sintaktikoa egiteko (Alegria et al., 2001). Hori gabe, luze joko luke honelako tresna bat garatzeak.

2. Euskarazko terminoak: ezaugarriak

Esan bezala, teknika linguistikoen lehen egitekoa da terminoen egitura morfosintaktikoak zehaztea. Euskarazko terminoen egitura aztertzen duten zenbait lan argitaratu dira (Elosegi, 314-338; Ensun-

za et al., 228-274; Perez Gaztelu, 1165-1181). Nolanahi ere, berariazko saioa egin dugu lan horietan zehaztutako egiturak lagin errealetan egiaztatzeko, eta, hala balegokio, egitura berriak edo zehatzagoak definitzeko.

2.1. Terminoen egitura

Lan honen aurretik, IXA taldeak ikerketa bat egin zuen terminoen egituraren inguruan, haien identifikazio automatikoari begira (Urizar et al., 2000). Hiru hiztegi terminologiko aukeratu ziren, eta horietako bakoitzetik 150 termino hautatu ziren, ausaz.

Lortutako emaitzetan, terminoen % 42 hitz bakarrekoak ziren (horien artean, % 70 izenak, % 23,6 aditzak eta % 6,4 adjektiboak). Bestalde, terminoen % 78,2 izen-sintagmak ziren, aditz-sintagmak % 18,2 eta % 3,4 adjektiboak. Hitz anitzeko terminoetan, izen-sintagmen proportzioa are handiagoa zen (% 83,2). Argi ikusi zen orduan hiztegien emaitzak testu errealen emaitzekin alderatu beharko zirela, sistema automatiko bat eratzeko benetan baliagarriak izango baziren.

Horretarako, corpus batetik terminoak eskuz erauzi ditugu. *zientzia.net*-eko informatika-alorreko 38 artikuluz osatua da erreferentzia-corpusa (42.302 hitz). Lan horren helburua bikoitza izan da: terminoen egitura aztertzeko bidea izateaz gain, tresna automatikoaren zehaztasuna ebaluatzeko ere erabili da. Hautaketa sistematikoa izan da; horretarako, testuak zoriz hautatu ziren, baina terminoen gehiegizko sakabanatzea saihesteko, testuak gutxienez 500 hitz izan behar zuen.

Hala ere, *Hapax Legomena* izeneko fenomenoa (hau da, terminoa behin bakarrik izatea testuan) nahi baino sarriago gertatu da. Hitz bakarreko terminoen % 28 inguru behin bakarrik agertu dira, eta hitz anitzeko terminoetan, ehuneko hori % 77ra iristen da. Horren arrazoia testuen izaera bera da; dibulgazio-lan motzak izaki, ez da harrigarria terminoak sakabanaturik agertzea. Gerora, erreferentzia-corpu-

sa testu luzeagoz osatu behar litzateke; ohartaraztekoa da, horrenbestez, oraingoan landu dugun lagina proba-banku interesgarri bezain gogorra dela.

Beste arazo nabarmen bat terminoak eskuz hautatzeko irizpideak finkatzea izan da, denok asetzeko moduko termino-definizioa aurkitzea zaila baita oso. Oinarrizko irizpideak finkatuta, hiru hizkuntzalarik markatu dituzte terminoak, termino luzeenaren irizpideari jarraiki betiere (hau da, termino batean beste bat habiatua dagoenean, luzeena aukeratu da beti). Adibidez, *posta elektronikoko mezu* terminoa agertu bada, hori bera hautatu da, horren barnean *posta elektroniko* termino laburragoa egonda ere. Hiru hizkuntzalarien arteko desbideratzeak aztertu dira, eta, erauzi beharreko terminoa adostu da. Horrek guztiak irizpideak areago fintzen lagundu digu.

Terminoa markatzeaz gain, haren egitura morfosintaktikoa ere zehaztu dugu. Markatutako termino bakoitzari dagokion eredua esleitu zaio. Ereduen osagaiak izendatzeko, nomenklatura hau erabili dugu:

- N izena
- N_{nc} deklinazio-kasurik gabeko (*non-case*)
- A adjektiboa
- A_{prep} izenlaguna
- A_{pos} izenondoa
- V aditza
- V_{gen} aditz-partizipioa genitiboarekin
- Adb adberbioa
- B bestelakoak

Bestalde, termino-osagai batzuk zein motatakoak diren erabakitzea ez da beti samurra; batik bat, zenbakiak, zenbaki-letra konbinazioak, akronimoak edo beste hizkuntzetatik zuzenean mailegatutako hitzak. Hona hemen adibide batzuk, bakoitzari esleitu zaion ereduaz hornituta: *Windows 98* ($N_{nc}N$), *Flip/flop* ($N_{nc}N$), *Pentium II* ($N_{nc}N$),

IBMren Aptiva ($A_{prep}N$), *AMDren k6 txip* ($A_{prep}N_{nc}N$). Urrats hau ondo egitea funtsezkoa da, eskuz esleitutako ereduek eta etiketatzaile automatikoak ematen dituenek bateragarriak izan behar baitute.

Erauzketa eskuz egiteko, XML editore bat erabili da, terminoaren testu-forma, forma kanonikoa (edo lema) eta eredua etiketatu ahal izateko (dokumentuan bertan txertaturik).

Aurreko urrats guztiak egin ondoren, erreferentzia-corpuseko termino-ereduen maiztasunak kalkulatzeko moduan gaude. Taula honetan bildu ditugu:

Termino-mota	**Eredua**	**Maiztasuna**	**Estaldura** (%)		
Bakunak	N	983	31,21		
	V/Adj/Adv	168	5,33	**Hitz anitzekoetan** (%)	
Hitz anitzekoak	$N_{nc}N$/FF	617	19,59	30,87	83,14
	$N_{nc}A_{pos}$	343	10,89	17,16	
	$A_{prep}N$	308	9,78	15,41	
	$N_{nc}A_{prep}N$	102	3,24	5,10	
	$N_{nc}N_{nc}N$/FFF	87	2,76	4,35	
	$A_{prep}N_{nc}N$	51	1,62	2,55	
	$N_{nc}N_{nc}A_{pos}$	30	0,95	1,50	
	$N_{nc}A_{prep}A_{pos}$	30	0,95	1,50	
	$A_{prep}N_{nc}A_{pos}$	26	0,83	1,30	
	$N_{abs}V_{gen}N$	19	0,60	0,95	
	$N_{nc}N_{nc}N_{nc}N$/FFFF	15	0,48	0,75	
	$N_{nc}A_{prep}N_{nc}N$	12	0,38	0,60	
	$A_{prep}A_{prep}N$	11	0,35	0,55	
	$N_{nc}N_{nc}A_{prep}N$	11	0,35	0,55	
	Bestelako izen-sint.	225	7,14	11,26	16,86
	Bestelako ereduak	112	3,56	5,60	
	Guztira	3.150	100	100	100

1. taula. Eskuz erauzitako terminoen ereduak

Izen-sintagmen nagusitasuna, espero bezala, erabatekoa da, terminoen % 87 inguru horrelakoak baitira. Bestetik, hitz anitzekoak gehiago dira (% 83,14) hitz bakarrekoak baino (% 31,21). Oro har, ondorio horiek berretsi egiten dituzte IXA taldeak aurreko azterketan lortutako emaitzak, baina corpus baten hustuketatik atera ditugunez, fidagarriagotzat jo behar dira.

2.2. Terminoen aldaerak

Terminoen aldaerak izan dira gure bigarren aztergaia. Terminologiaren alorreko hainbat adituk ohartarazi duenez, hiztegi terminologiko zein teknikoetako terminoak eta testuetan erabiltzen diren formak ez dira beti bat, eta, horregatik, terminoen aldaerak aztertzea funtsezko helburua da *Terminologiaren teoria komunikatibo* izeneko proposamen berrian (Cabré, 2001). Horrek eragin handia du ez bakarrik informazioaren berreskurapenaren alde praktikotik, terminologoen beharren aldetik ere bai. Terminoen erabilera deskribatu nahi denean, testuinguruan oinarritzea funtsezkoa da tresna erabilgarriak egin nahi badira, eta, horretarako, aldaeren detekzioa eginbehar nagusietako bat da. Are gehiago, ikuspuntu arautzailetik ere ezinbestekoa da jakitea zein diren kontzeptuak adierazteko aldaera desberdinak. Gure kasuan, terminologia erauzteko tresna baten ezaugarri erakargarria da terminoen aldaerak erlazionaturik aurkeztu ahal izatea.

Lehen arazoa kontzeptua da, *termino-aldaera* delakoaren definizioa bera lausoa baita, eta fenomeno anitz barneratzen baititu. Batetik, aldaerak, *sensu stricto*, kontzeptu beraren adierazle diren formak dira. Hala ere, hainbat autorek kontzeptu baten zehaztapenak (hiponimoak) edo hedapenak adierazten dituzten formak edo beste termino batekin eratzen dituzten konbinazioak (juntadura) aztertzen dituzte aldaerez dihardutenean.

Terminoak erauzteko proiektuetan, termino-aldaeren ikerketaren abiaburuan *termino kontrolatuak* edo *oinarri-terminoak* izaten dira.

Hainbat metodo linguistikotan (Daille, 1995; Jacquemin, 2001), erregelak definitu dira termino kontrolatuak legozkiekeen aldaerekin erlazionatzeko. Metodo estatistiko hutsetan, elkartze-neurriak aplikatu ohi dira lehenik biko terminoak lortzeko, eta gero horien aldaerak izan litezkeen termino luzeagoak bilatzen dira (Dias, 2). Batzuetan, termino berriak ere bilatzen dituzte (Lapata et al., 2003).

Aztergai horri ekiteko, euskarazko terminoetan ageri den aldakortasuna aztertu dugu.

2.2.1. Aldaeren sailkapena

Bibliografian proposatu diren aldaera-tipologiak ez datoz guztiz bat. Eragin handiena izan dutenak aintzat harturik (Jacquemin, 143; Daille, 9-1), eta euskararen ezaugarriak kontuan harturik, honako aldaeramotak bereizi ditugu: ortotipografikoak, morfologikoak, morfosintaktikoak, sintaktikoak eta semantikoak. Ikusiko dugunez, aldaera morfologiko eta morfosintaktiko batzuen arteko aldea lausoa da.

2.2.1.1. Aldaera ortotipografikoak

Aldaera tipografiko nagusiak dira letra larria/xehea aukera eta karaktere tipografiko batzuen erabilera (marra batez ere). 2. taulan adibide batzuk azaltzen dira. Bestetik aldaera ortografikoak multzo honetan dira sailkatzekoak.

Aldaera	**Adibideak**
letra larri/xehea	*Informatika Sail / informatika sail* *Internet / internet*
marratxoaren erabilera	*programazio lengoaia / programazio-lengoaia*
a itsatsia	*hizkuntza-tresna / hizkuntz tresna* *telefonia-sare / telefoni sare /*
lema-aldaerak	*iharduera / jarduera* *webgune / web gune*

2. taula. Zenbait aldaera ortotipografiko

IXA taldearen EUSLEM lematizatzaile-etiketatzailea (Ezeiza et al., 1998) gai da azken bi aldaera-motak identifikatzeko, eta forma normalizatua itzultzen du. Beraz, aldaera ortografikoak forma estandarraren bidez normalizatzen dira. Horien artean, batetik, termino bakunen arteko normalizazioa dago, aldaera estandarra eta ez-estandarra biak bakunak direnean. Adibidez, *iharduera/jarduera* edota *zientzilari/zientzialari* bikoteak aldaera estandarraren bidez normalizatzen dira (hau da, *jarduera* eta *zientzialari*). Bestetik, *liburudenda/liburu-denda* edota *webgune/web gune* erako aldaera-motak ere badaude (loturik/bereiz aldaera-motak, marraren idazkera ere tartean izaten dela).

2.2.1.2. Aldaera morfologikoak

Zenbait erdaratan, izenen singular/plural eta aditzen infinitibo/partizipio/gerundio aldaerak aipatu ohi dira (Jacquemin et al., 2000; Nenadic et al., 2002). Euskararen kasuan, flexioaren prozesatzea eginkizun nagusia dugu. Izan ere, lema bat edo, hitz anitzeko terminoen kasuan, forma kanoniko bat hainbat eratara flexionaturik ager daiteke testuan. Adibidez, *sistema eragile* terminoaren hainbat testu-forma aurkitzen dira: *sistema eragilearen*, *sistema eragileari, sistema eragileetan*, *sistema eragiletan*. Aldaera edo flexio horiek ez dira esanguratsuak terminoa identifikatzeko, hau da, termino *beraren* forma flexionatuak dira, baina flexio-aldaerak esanguratsuak izan daitezke terminoaren osagaien artean agertzen badira. Adibidez, *Interneteko konexio* eta *Interneterako konexio* formak termino-aldaeratzat dira hartzekoak; hala ere, flexioa barnean gertatzeak ez dakar beti berekin aldaerak izatea; esaterako, *sistemaren eragile* forma ez da, inolaz ere, *sistema eragile* terminoaren aldaera. Dena den, argitu nahi bada termino barneko flexio-aldaerak terminologiari begira noiz hartu behar diren kontuan, testuinguruari begiratu behar zaio, eta horregatik, uste dugu hurrengo atalean aztertuko diren aldaera morfosintaktikoekin batera lantzekoak direla.

Bestalde, tratamendu automatikoari begira, gure esku dago morfologiaren prozesamendua zertara bideratu, aipatu ditugun IXA tal-

dearen tresnei esker. Tresna horien bidez, aukera dugu forma kanoniko baten forma flexionatu guztiak lortzeko, edo forma flexionatu baten forma kanonikoa lortzeko. Beraz, bi osagaiko terminoetan, posible da flexio-aldaerek eragindako aldakuntzak batzea, dagozkien *lema-lema* edo *forma-lema* bikoteak hautatuz.

Flexio-aldaerez gain, eratortze-atzizkien bidezko aldaerak ere sartzekoak dira sail honetan; esaterako, *ordenagailu / ordenadore, murriztketa / murriztapen, kudeaketa / kudeatze*...

2.2.1.3. Aldaera morfosintaktikoak

Mota honetan, aurreko atalean aurkeztu diren zenbait aldaeraz gain, hitzen arteko ordena-aldaketak eta osagaien kategoria-aldaketak eragindakoak sailkatu ohi dira. Kontuan hartu diren osagai nagusiak hauek dira: izenak (N), izenlagunak (A_{prep}), izenondoak (A_{pos}), eta aditz nominalizatuak (N_v). 3. taulan azaltzen dira maiztasun handieneko baliokidetzak.

Aldaera morfosintaktikoak

Mota	**Eredua**	**Adibideak**
izen↔izenlagun	$N_{nc}N \leftrightarrow A_{prep}N$	*sare-kudeaketa ↔ sarearen kudeaketa*
	$N_{nc}N_{nc}N \leftrightarrow NA_{prep}N$	*Euskal Filologia Departamentua ↔ Euskal Filologiako Departamentua* *datu-base sistema↔ datu-baseen sistema*
	$N_{nc}N_{nc}N \leftrightarrow A_{prep}N_{nc}N$	*Excel kalkulu-orri ↔ Exceleko kalkulu-orri*
	$N_{nc}A_{prep}N \leftrightarrow A_{prep}A_{prep}N$	*barne datuen bus ↔ barneko datuen bus*
	$A_{prep}N_{nc}N \leftrightarrow N_{nc}A_{prep}N$	*Interneteko zerbitzu-hornitzaile ↔ Internet zerbitzuen hornitzaile* *mikroprozesadorearen barne-diseinu ↔ mikroprozesadore barneko diseinu*
izen↔aditz-izen	$N_{nc}N \leftrightarrow NN_v$	*sare-kudeaketa↔ sarea(k) kudeatze*
izen↔adjektibo	$N_{nc}N \leftrightarrow N_{nc}A_{pos}$	*informatika-ekipo↔ ekipo informatiko*

3. taula. Aldaera morfosintaktikoen tipologia

Hirugarren adibidean, ordena-aldaketa gertatzen da, baina aldaketa morfologikoak ere bai. Aldaera-mota hori konplexu samarra da, eta bi mota hartzen ditu bere baitan, morfologikoa eta sintaktikoa.

2.2.1.4. Aldaera sintaktikoak

Hauek dira bereizi ohi diren mota nagusiak:

- Txertaketa: oinarri-terminoaren osagaien tartean osagai berria txertatzen da (*telefonia-sare* ↔ *telefonia finkoko sare*)
- Osaketa (*overcomposition*): bi mota bereizi izan dira (Daille, 1995)
  - Ordezkapena (*Internet zerbitzu* + *zerbitzu-hornitzaile* ↔ *Interneteko zerbitzu-hornitzaile*
  - Alborakuntza[1] (*datu-base* ↔ *datu-base lexikal*)
- Juntadura: sintagmen juntaduraren emaitza: *AB eta C* sintagma, adibidez, *AC* terminoaren juntadurazko aldaera litzateke. Osagai komuna burua izan daiteke (*hizkuntza-tresnak eta -baliabideak*) edo modifikatzailea (*ekonomia- eta zuzenbide-ikasketak*)

Lehen biak flexioaren aldaerekin konbina daitezke. Hainbat adibide ageri dira 4. taulan (adibideak gure erreferentzia-corpusetik atera ditugu; testuan idazkera ez-estandarrean zeudenak idazkera estandarrera aldatu ditugu taulan).

[1] Jacquemin-ek (2001) ez ditu horrelakoak aldaeratzat hartzen, haren ustez ez dutelako ezinbesteko baldintza bat betetzen: aldaerak ez luke jatorrizkoaren barruan egon behar. Daille-k (1995, 2001), berriz, mota honetan sailkatzen ditu.

Aldaera sintaktikoak

Mota		Eredua	Adibideak
Txertaketa	Determinatzailea	$A_{prep}N \leftrightarrow A_{prep}\mathbf{D}N$	*audioko formatu ↔ audioko azken formatu*
		$N_{nc}A_{prep}N \leftrightarrow N_{nc}A_{prep}\mathbf{D}N$	*posta elektronikoko mezu ↔ posta elektronikoko zenbait mezu*
		$N_{abs}V_{gen}N \leftrightarrow N_{nc}\mathbf{D}_{abs}V_{gen}N$	*informazioa bilatzeko sistema ↔ informazio hori bilatzeko sistema*
		$N_{abs}V_{gen}N \leftrightarrow N_{nc}\mathbf{A_{pos}}V_{gen}N$	*informazioa bilatzeko sistema ↔ informazio fidagarria bilatzeko sistema*
		$NAdv \leftrightarrow N\mathbf{D}Adv$	*gigabit segundoko ↔ gigabit bat segundoko*
	Izenlaguna	$N_{nc}N \leftrightarrow N_{nc}\mathbf{A_{prep}}N$	*telefonia-sare ↔ telefonia finkoko sare*
		$A_{prep}N \leftrightarrow A_{prep}\mathbf{A_{prep}}N$	*barneko bus ↔ barneko datuen bus*
	Adjektiboa	$N_{nc}A_{prep} \leftrightarrow N_{nc}\mathbf{A_{pos}}A_{pos}$	*batuketa haztatu ↔ batuketa aljebraiko haztatu*
	Izena	$N_{nc}N \leftrightarrow N_{nc}\mathbf{N_{nc}}N$	*386 ordenadore ↔ 386 SX ordenadore*
		$N_{nc}N_{nc}N \leftrightarrow N_{nc}N_{nc}\mathbf{N_{nc}}N$	*486 SX mikro ↔ 486 SX 25 mikro*
		$NabsV_{gen}N \leftrightarrow NabsV_{gen}\mathbf{N_{nc}}N$	*idazkera ezagutzeko programa ↔ idazkera ezagutzeko ordenadore-programa*
	Postposizioa	$N_{nc}N \leftrightarrow A_{prep}\mathbf{A_{prep}}N$	*neurona-sinapsi ↔ neuronen arteko sinapsi*
		$N_{nc}N + N \leftrightarrow N_{nc}A_{prep}\mathbf{A_{prep}}N$	*memoria-banku + trukaketa ↔ memoria-bankuen arteko trukaketa*
Osaketa	Ordezkapena	$N_{nc}N + N_{nc}N \leftrightarrow A_{prep}N_{nc}N$	*Internet-zerbitzu + zerbitzu-hornitzaile ↔ Interneteko zerbitzu-hornitzaile*
	Alborakuntza	$N_{nc}N \leftrightarrow N_{nc}N_{nc}\mathbf{N}$	*bezero/zerbitzari ↔ bezero/zerbitzari arkitektura*
		$N_{nc}A_{prep} \leftrightarrow N_{nc}A_{pos}\mathbf{A_{pos}}$	*batuketa aljebraiko ↔ batuketa aljebraiko haztatu*
		$N_{nc}N_{nc}A_{pos} \leftrightarrow N_{nc}N_{nc}A_{pso}\mathbf{A_{pos}}$	*harpidedun-linea digital ↔ harpidedun-linea digital asimetriko*

Mota			Eredua	Adibideak
Juntadura	Bigramak	Burua	$N_{nc}N \leftrightarrow N_{nc}\mathbf{NJN}$	*programazio-lengoaia + programazio metodoak ↔ programazio-lengoaiak eta -metodoak*
			$A_{prep}N \leftrightarrow A_{prep}\mathbf{NJN}$	*ordenagailuetako hardware + ordenagailuetako software ↔ ordenagailuetako hardware eta software*
			$N_{nc}A_{prep} \leftrightarrow \mathbf{N_{nc}JN_{nc}}A_{pos}$	*lengoaia informatikoak + sistema informatikoak ↔ lengoaia eta sistema informatikoak*
		Argumentua	$N_{nc}N \leftrightarrow \mathbf{N_{nc}JN_{nc}}N$	*irrati-emanaldiak + telebista-emanaldiak ↔ irrati- eta telebista-emanaldiak*
			$A_{prep}N \leftrightarrow \mathbf{A_{prep}JA_{prep}}N$	*ortografiaren zuzentzaile + estiloaren zuzentzaile ↔ ortografiaren eta estiloaren zuzentzaile*
			$A_{prep}N + A_{prep}N \leftrightarrow \mathbf{N_{nc}JA_{prep}}N$	*bideoaren kontrolatzaile + diskoaren kontrolatzaile ↔ bideo eta diskoaren kontrolatzaile*
			$N_{nc}A_{pos} \leftrightarrow N_{nc}\mathbf{A_{pos}JA_{pos}}$	*arau sintaktikoak + arau semantikoak ↔ arau sintaktiko eta semantikoak*

4. taula. Aldaera sintaktikoak

2.2.1.5. Aldaera semantikoak

Aldaera semantikoen funtsa terminoen osagaien arteko sinonimia-erlazioak dira (esan gabe doa, morfologiaren bidez ezin azal daitezkeen sinonimia-erlazioak). Adibidez, *idazkera bitar* eta *notazio bitar* edo *hizkuntz atlas* eta *atlas linguistiko* terminoak baliokideak dira. Erauzlearen bertsio honetan ez ditugu landu, oraingoz ez baitugu honelako erlazioak modu sistematiko eta zehatzean tratatzeko modurik. Zalantzarik gabe, hurrengo bertsioan aurre egin beharreko egitekoa da.

3. Lan esperimentala

Prozesatze automatikoa bi urrats nagusik osatzen dute: a) hautagaiak teknika linguistikoen bidez atzematea; eta b) teknika estatistikoak erabiliz iragaztea eta ordenatzea. Gainera, sistemaren etekina hobetzeko, termino habiatuak ere hautagaitzat hartu behar dira.

Funtsezko prozesu horiez gain, garatu dugun tresna informatikoak eskaintzen du, batetik, corpusak eratzeko modulu bat, eta, bestetik, erauzketaren emaitzak bistaratzeko eta kudeatzeko interfaze grafiko bat. Arkitektura informatikoari begira, beraz, hiru osagai nagusi daude: corpusaren eraikitzailea, terminoen etiketatzailea eta corpus etiketatuaren gaineko nabigatzailea. Etiketatzeko, XML metalengoaia erabili da; corpusak eta erauzketaren emaitzak biltegiratzeko eta kudeatzeko, XML datu-base bat (Berkeley DB XML).

Azkenik, esan behar da ebaluazio-tresna berezi bat eratu dela, sistema automatikoaren doitasuna eta estaldura modu automatikoan kalkulatu ahal izateko.

Estaldura eta doitasuna dira *Informazioaren Berreskurapena eta Erauzketa* (ingelesezko IR eta IE laburtzapenak erabili ohi dira) izeneko arloetan erabiltzen diren ebaluazio-neurri estandarrak. Doitasu-

nak (*Precision* - P) erauzitako informazioaren kalitatea neurtzen du, hau da, gure kasuan, erauzitako terminoetatik zenbat diren benetako termino. Estaldurak (*Recall* - R), berriz, erauzi beharko liratekeenetatik zenbat erauzi diren neurtzen du, hau da, testuan dauden terminoetatik, programak zenbat erauzi dituen.

Estaldura eta doitasun handiak batera lortzea oso zaila da; beraz, sistema erabat automatikoetan doitasunari ematen zaio lehentasuna, eta erdiautomatikoetan, berriz, estaldurari. Hain zuzen ere, bigarren hori da gure kasua, hau da, ahalik eta estaldura handiena ziurtatu nahi dugu. Izan ere, erabiltzaileari emaitza eskuz lantzeko (hau da, benetako terminoak hautatzeko) aukera emango zaio, eta, horretara, doitasuna handitzeko aukera.

Doitasuna eta estaldura bateratzen duen *F neurri* izeneko neurri bat dago (*F*). Doitasunak eta estaldurak bat egiten duten puntua kalkulatzen du, formula honen bitartez: $F = 2PR/(P+R)$

Hurrengo ataletan, garatutako tresnaren hiru atal nagusiak azalduko dira; lehen biak, prozesatze linguistikoa eta prozesatze estatistikoa, terminoen etiketatzailearen parte dira; hirugarrena, berriz, erabiltzaileari begirako aplikazioa da, corpusaren eraikitzailea eta nabigatzailea integratzen dituena.

3.1. Prozesatze linguistikoa

Prozesatze linguistikoaren emaitzan estaldura maximoa lortzeko asmoz, urrats hauek egin dira:

- terminoen egitura morfosintaktikoa identifikatzen duen gramatika idaztea
- gramatikaren hobekuntzak
- normalizazioa eta terminoen habiatuen tratamendua

3.1.1. Gramatikaren definizioa

Prozesatzeko azkarra eta idazteko nahiko sinplea den egoera finituko gramatika bat idatzi da, *xfst* teknologia erabiliz (Beesley et al., 2004). Osagaien oinarrizko informazioa erabilitako lematizatzaile/etiketatzailearen araberakoa izan da, eta, gramatika idazterakoan, 1. taulan aztertutako eredu morfosintaktikoak hartu dira kontuan.

Gramatika egiteko, IXA taldearen aurreko gramatika batetik abiatu gara, eta terminologia-erauzketaren helburu eta ezaugarrietarako moldatu eta garatu dugu. Mapatze bat egin behar izan da eskuz landutako ereduen eta lematizatzaile/etiketatzailearen irteeraren artean. 5. taulan azaltzen dira mapatze horretan erabilitako baliokidetza nagusiak.

Eredua	**Gramatika**	**Kategoria**	**Azpikategoria**	**Kasua**
N	N_o | N_{nc} | N_{prep}	IZE | SIG | IDENT	ARR | IZB | LIB	DEK |ABS | GEN | GEL
A_{prep}	A_{prep}	N | ADJ | ADI	IZE | IZO | IZL	GEN | GEL | BAN | DESK
A_{pos}	A_{pos}	ADJ | ADI	IZO	

5. taula. Gramatikaren eta etiketatzailearen etiketen arteko mapatzearen laburpena

Eskuz egindako erauzketan oinarrituta, oso erraza da gramatika horrekin lortzen den estaldura kalkulatzea. Gramatikan ez daude eredu guztiak, baina bai maiztasun handienekoak, hortxe baitago koska, alegia, zenbat eta eredu gehiago sartu, hainbat eta estaldura handiagoa lortzen da, baina doitasuna ere jaitsi egiten da; beraz, ondo haztatu behar da zein eredu sartu gramatikan eta zein ez.

Oraingoz, izen-sintagmen gramatika bat definitu da, hitz bakarreko zein hitz anitzeko terminoen eredu nagusiak jasotzen dituena. 1. taulan azaldu denez, izen-sintagmak terminoen % 90 dira kasik; bestalde, aditz-sintagmak mugatzea konplexu samarra da, ordena ez da

beti finkoa eta, gainera, izen-sintagmak txertatu ohi dira aditz-sintagmaren osagaien artean.

Lehen urratsean, izen-sintagma luzeenak erauzten dira. Sintagma luzeen barnean dauden eta ereduekin bat datozen azpisintagma batzuk termino-hautagaiak izan daitezke. Baina horretarako baldintza bereziak bete behar dituzte, izen-sintagma barneko hitz-segida guztiak ez baitira azpisintagmak, eta aintzat hartuz gero, zarata handia sortuko lukete. Horien tratamendua aurrerago azalduko dugu (**Termino habiatuak** atala).

Bestalde, esan bezala, maiztasun txikiko hitz anitzeko eredu batzuk ez dira sartu gramatikan. Jarraitutako irizpidea eskuzko corpusean duten maiztasuna izan da, behin bakarrik edo oso gutxitan agertzen diren ereduak ez baitziren 'errentagarritzat' jo, behar den prozesatze-lana kontuan harturik. Maiztasun txikiko zenbait eredu landu ditugu; horretarako arrazoia izan da 'emankorrak' iritzi diegula (estaldura aldetik); hau da, gure irudipena da, erreferentzia-corpusean urriak izan arren, horrelako ereduen araberako terminoak ez direla euskaraz hain bitxiak. Adibidez, *telekomunikazioen munduko enpresa* adibidetzat daukan $A_{prep}A_{prep}N$ eredua. Alderantziz, maiztasun handiagoko zenbait eredu ez dira gramatikan landu, zarata sortzeko arriskua handiegia izan zitekeelakoan ($N_{nc}N_{nc}N_{nc}N$, $N_{abs}V_{gen}N$ eta $N_{nc}N_{nc}A_{prep}N$). Adberbio gehienak ere baztertu dira, zarata handia sortzen dutelako, eta termino gutxi identifikatzeko gai direlako. Hitz bakarreko terminoetarako, izenak, akronimoak (SIG) eta karaktere-konbinazio berezi batzuk (IDENT) dira onartutako mapatzeak.

Gramatika definitu ondoren, erauzketa automatikoan espero dezakegun estaldura-balio handiena % 84,69 da, hori baita gramatikan sartu diren ereduei dagozkien terminoen ehunekoa esku-erauzketaren emaitzetan.

3.1.2. Oinarrizko prozesuaren ebaluazioa eta hobekuntza

Gramatikaren lehen bertsioaren bidez, terminoen % 67 baino ez ziren identifikatu. Aipatutako eredu-murrizketaren ondorioz jaitsiera bat espero zitekeen, baina ez hainbestekoa. Arazoa aztertu eta erroreen iturburua identifikatzen hasi ginen:

- Arazo tipografikoak. Terminoen osagai batzuk, zifradun identifikadoreak adibidez, ez ziren etiketatzen, haien ezaugarri bereziak zirela kausa. Marratxoak eta barra etzanak ere arazoak ematen zituzten (adib. *flip/flop*, *software-teknologia* edo *TCP/IP protokoloa*). Marratxoaren kasuan, arazoa begi-bistakoa zen: gramatikak bi izen bilatzen zituen, baina etiketatzaileak izen bakartzat hartzen zituen
- Lematizatzaile/etiketatzailearen erroreak. Tresnak egiten duen desanbiguazio morfosintaktikoa ez da akatsik gabea, eta batez beste errore gutxi egiten badu ere, % 5 inguru, darabilen lexikoan ez dauden hitzen analisietan metatzen dira errore gehienak, eta terminoetan horrelako asko dago
- Erdaretako terminoak edo mailegu irentsi gabeak, *HyperText Markup Language* esaterako
- Normalizazioa eta termino habiatuen arazoa (hurrengo ataletan azalduko ditugu)

Arazo horiek konpondu ahal izateko, lehen urrats batean bi aldaketa egin ditugu, bata corpusetik independentea, eta bestea corpusaren mendekoa, hau da, gainbegiratua.

Lehen bidetik, sinbolo berezien identifikazioa berariaz landu da, eta aurreko lehen arazoari irtenbidea eman zaio.

Bigarren eta hirugarren arazoari aurre egin nahian, erabiltzailearen hiztegi bat gehitu zaio sistemari; horri esker, lematizatzaile/etiketatzaileak, bere lexiko orokorraz gain, lexiko berezi bat ere kudeatzeko

gai da. Horrek corpusa aurrez aztertzea eskatzen du, eta lan hori eskuz edo modu erdiautomatikoan egin daiteke. *Xuxen* zuzentzaile ortografikoa (Aduriz et al., 1996) eskura izaki, modu erdiautomatikoan egin ahal izateko funtzionalitatea barneratu da *Erauzterm*-en.

Tresna hori erabiliz, lematizatzaile/etiketatzaileak identifikatzen ez dituen testu-hitzak maiztasunaren arabera ordenatzeko aukera dago eta, erabiltzailearen interesen arabera, halako testu-hitz kopuru bat hautatzeko eta lantzeko ere bai (lema eta kategoria esleitzea). Horrela landutako lemak erabiltzailearen hiztegira esporta daitezke. Arazoa ez da erabat konpontzen, baina, kostu-etekin erlazioari erreparatuz gero, hobekuntza ukaezina da. Beraz, aztertzekoa litzateke ea merezi lukeen *Erauzterm*-en erabiltzaileari erauzketarako testuak horrela, modu erdiautomatikoan, aurreprozesatzeko aukera eskaintzea.

Esan bezala, erabiltzailearen hiztegiak ez ditu konpontzen etiketatzaileak hiztegiko lemekin egiten dituen errore guztiak. Horren aurrean, etiketatzailea hobetzea da bide zuzenena, baina, bitartean, gramatikaren bitarteko zenbait konponketa egin ditugu. Ikusi dugu, adibidez, EUSLEMek ezagutzen ez dituen testu-hitzen kasuan, lexikorik gabeko lematizazioa zuzena izan arren, kategoria esleitzean erroreak gertatzen direla batzuetan (adibidez, izen direnei ADJ kategoria ematea); datuak aztertuta, eta espero izatekoa zenez, horrelako gehien-gehienak izenak dira berez, eta lexikorik gabeko lematizazioan proposatzen diren lemak izentzat hartzea lehenetsi dugu. Kategoria esleitzea dela eta, testuingurua ere hartu dugu kontuan; esaterako A_{prep} baten ondorengo testu-hitzari EUSLEMek ADJ kategoria esleitu dionean, egiaztatu dugu ia beti IZE kategoriakoak direla (adibidez, *-tzaile* atzizkidunekin gertatu ohi da hori: *sistemaren kudeatzaile*); hori dela eta, erabaki dugu ereduetako N elementuarekin mapatzea A_{prep} baten ondorengo ADJ kategoria. Bestalde, formaz partizipio mugatuak direnak (*banatua, banatuak...*), eta etiketatzailearen irteeran aditz-partizipioaren analisia dutenak, A_{pos}-ekin parekatzea eraba-

ki dugu, aurretik N_{nc} baldin badute (*sistema banatua, hitz isolatuak...*). Horrelako beste zenbait konponketa egin dira. Horren guztiaren helburua estaldura handitzea da, baina aitortu behar da doitasuna zertxobait jaistea ekar dezakeela berekin.

3.1.3. Aldaeren tratamendua: normalizazioa eta aldaeren arteko erlazioa

Aldaeren tratamendua dela eta, aldaera-mota batzuk normalizatu egin dira, hau da, maiztasunak eta neurri estatistikoak kalkulatzean, aldaeren agerraldiak forma normalizatu edo 'kanoniko' bati esleitu zaizkio. Horrela landu dira, adibidez, aldaera ortografiko/tipografiko gehienak eta, zer esanik ez, termino-bukaerako flexioak (hau da, *forma*-lema* erakoak dira termino kanonikoak). Horretara, *Internet* eta *internet*-en agerraldiak batera zenbatu dira; orobat *hizkuntz eredu/hizkuntza-eredu/hizkuntza eredu*; *iharduera-eremu/jarduera eremu*, eta abar. Erauzketaren emaitzatzat eskaintzen den termino-zerrendan, aldaera horiei guztiei esleitutako forma kanonikoa soilik agertzen da (adibidez, *jarduera-eremu*); aldaerak bere horretan ageri dira forma kanonikoari dagozkion testuinguruetan. Horiek guztiak etiketatzaileak ebazten ditu zuzenean.

Gainerako aldaera-motak, berriz, beste era batera landu dira. Batetik, emaitzetan agerraldiak bereiz kontatu dira. Esaterako, *Interneteko konexio* eta *Interneterako konexio* bi termino dira, eta, beraz, bakoitzaren neurri estatistikoak kalkulatu dira, eta emaitzen zerrendan zeinek bere forma kanonikoa du. Baina, bestetik, horrelako aldaerak elkarren artean erlazionatzeko sistema garatu dugu. Horren bidez, jakin dezakegu halako termino baten aldaera detektatu dela corpusean, eta zuzenean bistaratu horri dagozkion informazioa eta emaitzak.

Aldaerak topatzeko sistema horren ezaugarri nagusia da terminohautagaia bi mailatan deskribatzea. Bata maila sintagmatikoa da, eta bestea, paradigmatikoa (Jacquemin, 1999). Maila sintagmatikoan,

egitura morfosintaktikoa definitzen da. Maila paradigmatikoan, aldiz, osagaien informazioa: lema eta erroa. Oraingoz ez dugu erroak ateratzeko baliabiderik; beraz, erregelak sortzeko tresnak onartzen dituen arren, erroak behar dituzten erregelak ez dira oraingoz aplikatu aldaeren tratamenduan.

sare-kudeaketa	maila sintagmatikoa	$\{N_0 \rightarrow N_{nc1}\ N_2\}$
	maila paradigmatikoa	$(N_{nc1}$ lema$)$=sare $(N_2$ lema$)$=kudeaketa

6. taula. S*are-kudeaketa* terminoaren egitura, maila sintagmatikoan eta paradigmatikoan deskribatuta

sarearen kudeaketa	maila sintagmatikoa	$\{N_1 \rightarrow A_{prep1}\ N_2\}$
	maila paradigmatikoa	$(A_{prep1}$ lema$)$=sare $(N_2$ lema$)$=kudeaketa

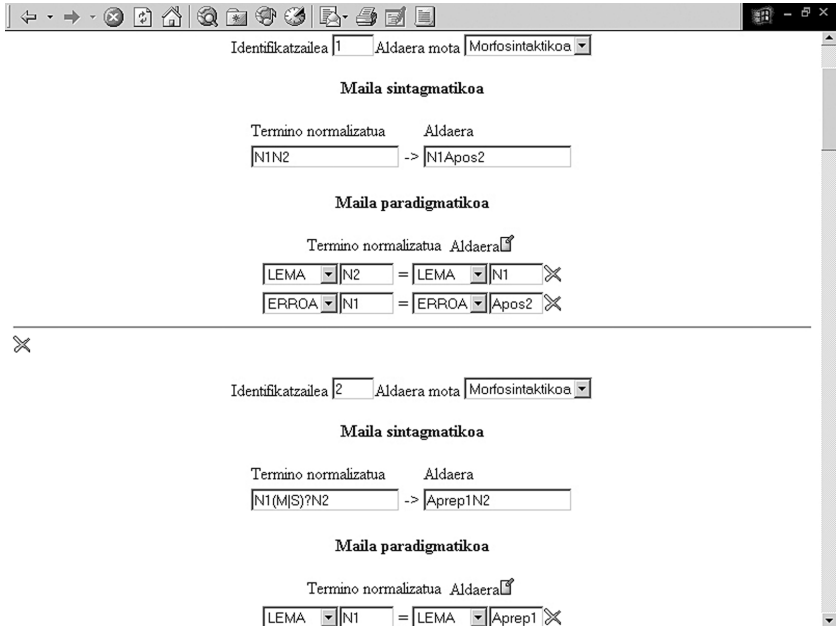
7. taula. S*arearen kudeaketa* aldaeraren egitura, maila sintagmatikoan eta paradigmatikoan deskribatuta

Maila horiek bereiziz, termino-hautagaien arteko maila bereko loturak egin ditzakegu. Loturen bidez, forma kanonikoa dagokion aldaerarekin erlaziona daiteke. Loturak definitzeko erregelak honako formatua du:

maila sintagmatikoa	$\{(N_0 \rightarrow N_{nc1}\ N_2) \Rightarrow (N_1 \rightarrow A_{prep1}\ N_2)\}$
maila paradigmatikoa	$(N_{nc1}$ lema$)=(A_{prep1}$ lema $)$ $(N_2$ lema$)=(N_2$ lema$)$

8. taula. $N_{nc}N \leftrightarrow A_{prep}N$ aldaerak tratatzeko erregela

Formatu horretako erregelak definitzeko tresna bat garatu dugu (1. irudia). Erregela bakoitzaren maila sintagmatiko eta paradigmatikoetako loturak finkatzeko aukera ematen du. Behin erregelak egin-da, termino-hautagaiak gordetzen diren datu-baseko kontsulta-sistemaren sintaxian konpilatzen dira. Termino habiatuen tratamendua egin eta gero aplikatzen zaizkie termino-hautagai guztiei. Emaitzak, hau da, loturak eta lotura-mota (aldaera-mota), datu-basean adierazten dira. Honako aldaera-motak inplementatu dira orain arte: ortotipografikoak, *izen* ↔ *izenlagun* erako aldaera morfosintaktikoak, eta, aldaera sintaktikoen artean, txertaketa, ordezkapena eta alborakuntza.



1. irudia. Aldaeren erregelak sortzeko tresnaren interfazea

3.1.4. Termino habiatuak

Lehen esan dugu transduktoreak izen-sintagma luzeenak atzematen dituela lehen urratsean. Bistan da, ordea, terminoak ez duela zertan izen-sintagma luzeena izan, horren barneko azpisintagma bat izan baitaiteke. Esaterako, *sistema eragile merke* izen-sintagma ez da termino, baina bai horren barnean den *sistema eragile*. Horrelako terminoei *habiatuak* esaten zaie (*nested terms*). Horrelako terminoak erauzi ezean, ezin dugu espero estaldura-datuak taxuzkoak izatea. Sintagma osoak soilik erauzita, datu estatistikoak sakabanatuagoak izango dira, *sistema eragile*-ren agerraldiak *sistema eragile merke* eta antzeko sintagma luzeagoetan banatzen direlako. Horrek ekar dezake sintagma luzeak, neurri estatistikoak esanguratsuak ez izateagatik, emaitzetatik bazterturik geratzea; eta, ondorioz, termino den azpisintagma ere begien bistatik galtzea. Beraz, termino habiatuen tratamendua ezin saihestuzko eginkizuna da termino-erauzketan.

Arazo honi aurre egiteko, izen-sintagma luzeenak azpisintagmatan banatu dira. Argi utzi nahi dugu izen-sintagma osoaren edozein azpikate ez dela azpisintagmatzat jotzen, eta gramatika baten araberakoak direnak soilik hartzen direla kontuan. Gramatikaren funtsa burua eta modifikatzaileak (*head* eta *modifier*) bereiztea da, LEXTER proiektuan egiten zen bidetik (Bourigault, 1994). Adibidez, prozesaketa horren bidez *RAM memoria handi* termino luzean, benetako terminoa ez dena baina gramatikaren bidez lortu dena, *RAM memoria* termino zehatza ere identifikatzen da, eta hautagaien zerrendan barneratzen.

Aurreko prozesu guztien bitartez, estaldura % 73,78tik % 83,94ra igo da, 9. taulan ikus daitekeenez. Doitasuna, berriz, % 4 inguru jaitsi da, % 33,24tik % 29,36ra, baina gogoan izan behar dugu urrats honen helburu nagusia estaldura handitzea dela, eta doitasuna hobetzea metodo estatistikoen esku geratuko da.

	terminoak	hautagaiak	benetakoak	estaldura	doitasuna	F neurria
Monogramak	983	2.424	857	87,18	35,35	50,31
Bigramak	1.258	3.727	1.068	84,90	28,66	42,85
3-4gramak	318	1.164	223	70,13	19,16	30,09
GUZTIRA	2.559	7.315	2.148	83,94	29,36	43,51

9. taula. Prozesatze linguistikoaren emaitzak,
termino habiatuen tratamenduaren ostean

Estaldura gehiago igotzea oso zaila da, etiketatzailearen ohiz kanpoko erroreak eta maiztasun gutxiko termino-eredu ohiz kanpokoak direlako estaldura handiagoa lortzeko eragozpen nagusiak.

Nolanahi ere, termino-erauzketaren alorreko estandarrak kontuan harturik, lortu dugun estaldura aski ona da, eta, hurrena, doitasuna igotzen lagunduko diguten metodo estatistikoak azalduko ditugu.[2]

3.2. Prozesatze estatistikoa

Prozesatze linguistikoaren bidez lortutako hautagaiak sailkatzea da urrats honen helburua. Sailkapena egin ondoren, bi aukera daude:

- sistema erabat automatikoetan, muga edo atalase bat jartzen zaio erabiltzaileak aukeratutako neurri estatistikoari, eta hortik gora dauden terminoak hautatzen dira, ordenaturik
- sistema erdiautomatikoetan, erabiltzaileari (terminologoa, itzultzailea, hizkuntzalaria, idazlea…) emaitzen zerrenda eskaini

[2] Doitasuna teknika linguistikoen bidez handitzeko asmoa ere badugu. Adibidez, termino habiatuak ateratzean, postposizioek, berariaz lantzen ez badira, zarata handia sortzen dute; *ordenagailuen arteko komunikazio* terminotik ez genuke biko azpisintagmarik atera behar, ezin baitira termino-hautagaiak izan (**ordenagailuen arte*); hori egin ahal izateko, ordea, postposizioak atzemateko sistema bat behar da. IXAk garatua du dagoeneko sistema bat horretarako (*Zatiak* moduluaren barnean), eta erauzlearen hurrengo hobekuntzetako bat izango da hori integratzea.

ohi zaio, neurri estatistiko jakin baten arabera ordenatuta, bere ustez termino direnak edo, duen helburuaren arabera, interesatzen zaizkionak hauta ditzan. Aurreko atalean aipatu berri dugun atalasea ere ezar daiteke, erabiltzaileak emaitzen zerrenda murriztea komeni zaiola uste badu

Teknika estatistikoen aplikazioa asko aldatzen da proiektu batetik bestera. Termino-erauzketan, hitz bakarreko eta hitz anitzeko terminoek dituzten ezaugarri desberdinak direla kausa, estrategia desberdinak bideratu izan dira. Hitz anitzekoen elkartze-neurriek *unithood* ezaugarria neurtzen dute, hau da, osagaiek 'unitate' bat zenbateraino osatzen duten. Hitz bakarrekoei antzemateko, berriz, *termhood* izeneko alderdia hartu ohi da kontuan (Kageura, 1996); horren arabera, unitate bakuna jakintza-alor edo erabilera-alor bateko unitate bereizgarria zenbateraino den haztatzen da.

Dena den, elkartze-neurrien emaitza egokia izan dadin, beti eskatzen da agertze-maiztasun minimo bat, behin bakarrik agertzen direnen erauzketa doia (*Hapax Legomena*) oso arazo zaila baita. Hori dela eta, egokiagoa da testu luzeekin lan egitea, baina hori ez da beti posible test-corpusetan; beraz, emaitzak ebaluatzean kontuan hartu beharko da alderdi hori. Muga hori alde onera hartzea ere badago; alegia, erauzlearen estaldura eta doitasuna ebaluatzean kondizio txar samarrak ezartzen baditugu, uste izatekoa da emaitzak hobeak izango direla erauzketa testu-bilduma handietatik egiten denean.

10. taulan, terminoen luzeraren eta maiztasunaren arteko korrelazioa azaltzen da.

Mota	**maiztasuna=1**	**maiztasuna=2**	**maiztasuna>2**
hitz bakarreko terminoak	% 28,20	% 16,00	% 55,80
hitz anitzeko terminoak	% 77,47	% 12,91	% 9,62

10. taula. Terminoen maiztasunak motaren arabera

3.2.1. Hitz anitzeko terminoak

Hitz anitzeko terminoen gaineko neurri estatistikoak dira bibliografian gehien landu direnak. Neurri gehienek bigramen gainean lan egiten dute, hau da, bi hitzeko terminoekin. Neurri horietan sakontzea ez da artikulu honen helburua, baina gaingiroki azalduko ditugu, behintzat.

Bi hitzeko unitateak (terminoak, kolokazioak...) erauzteko tekniketan erabiltzen diren neurriei *lexical association measures* (AM) edo 'hitzen elkartze-neurriak' izendapena ematen zaie maiz (Evert, 2001). AMen bidez, hitz-bikote edo bigrama-multzo batetik (w_1, w_2) korrelazio handia duten bikoteak identifikatzen dira. 'Korrelazio' diogunean, elkarrekin agertzeko 'joeraz 'ari gara, hau da, hitz bakoitzaren maiztasuna kontuan harturik, zoriz legokiekeen baino maiztasun handiagoz agertzea elkarren ondoan. Hitz-konbinazioa terminoa ez bada, bi hitzak elkarren ondoan gertatzea zorizko gertakaria da, hitz bakoitza bere probabilitatearen arabera agertzen da testuan, bestetik independenteki.

(*A,B*) bigrama jakin bakoitzerako, AMek kontingentzia-taula honetako balioak konputatzen dituzte, bakoitzak bere eredu estatistikoaren edo probabilistikoaren arabera:

	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	O_{11}	O_{12}
$w_1 \neq A$	O_{21}	O_{22}

Neurriaren balioa zenbat eta handiagoa den, hainbat eta aukera handiagoa dago konbinazioa termino izateko. Hauek dira neurri erabilienak:

- *Elkarrekiko informazioa* (*Mutual Information* - MI). Informazioaren Teoriako kontzeptua da, eta Church eta Hanks-ek erabili

zuten lehen aldiz corpusetatik kolokazioak ateratzeko (1990). Terminologiaren erauzketan asko erabili da

- *Egiantz-arrazoia (Log-likelihood ratio* - LR). Hipotesi-egiaztatzearen alorrekoa. Banaketa binomialean oinarritutako testa da. Dunning-ek (1994) proposatu zuen arlo honetan erabiltzea, eta emaitzak egokiak izan dira hainbat proiektutan
- *Ji karratua* (*Chi-square* - χ^2). Hipotesi-egiaztatzearen alorrekoa. Pearson-en χ^2 testak datu-multzo bat χ^2 banaketara zenbateraino hurbiltzen den egiaztatzeko erabiltzen da
- *t neurria* (*t-score*). Hipotesi-egiaztatzearen alorrekoa
- Dice-ren koefizientea. Antza neurtzeko oso erabilia. Smadjak proposatu zuen kolokazioak erauzteko (Smadja, 1995)

Maiztasuna ere sartu ohi da elkartze-neurri heuristikoen artean (Evert, 2004). Erabili ohi den beste neurri heuristiko bat MI^3 da.

Bigrametarako neurriak hiru osagai edo gehiago dituzten unitateei aplikatzea ez da berehalakoa. Talde batek (Dias et al., 2000) *elkarrekiko itxaropena* (*Mutual Expectation*-ME) izeneko neurri berri bat proposatu du, Dice-ren koefizientearen orokortzea dena.

Badira proposamen pragmatikoago batzuk; esaterako *C-NC-SNC-balioak* (Maynard and Ananiadou, 2000), testuinguruko informazioa zein semantika ere gehitzen dutenak. Nakagawa-k (2003) termino habiatuen eragina hartzen du kontuan eta, azkenik, Lapata eta Lascarides-ek (2003) semantikan eta ikasketa automatikoan oinarritutako metodo bat proposatzen dute.

3.2.1.1. Bigramen azterketa

Prozesatze linguistikoaren emaitzetan oinarrituta, bi hitzeko terminoen zenbait neurri estatistiko kalkulatu ditugu, eta eskuz etiketa-

tutako erreferentzia-corpusarekin konparatuta ebaluatu dira emaitzak. Lan honetan, goian azaldu ditugun neurriak erabili ditugu (maiztasunaz gain): MI, MI^3, LR, ji karratua, t neurria eta Dice-ren koefizientea.

8. taulan bildu ditugu emaitzak. Bertan agertzen diren balioak F neurriaren balio onena ematen dutenak dira, hau da, doitasun/estaldura erlazio onenekoak.

	Doitasuna	**Estaldura**	**F neurria**
MI	29,25	82,91	43,24
MI^3	29,25	82,91	43,24
LR	31,90	68,04	43,44
Ji karratua	28,84	84,18	42,96
t neurria	30,93	72,42	43,35
Dice	28,91	84,66	43,10

11. taula. Bigramen lehen emaitzak

Prozesatze linguistikoan lortutako emaitzekin konparatuta (9. taula), ez da F neurriaren hobekuntzarik nabari. Alegia, metodo estatistiko hauen ekarpena txiki samarra da.

Hala ere, hitzen maiztasuna neurtzean, erauzketa egiten ari den corpusean duten maiztasuna erabili da. Corpusen tamaina txikia dela kontuan hartuz, hitzek corpus orokor batean duten maiztasuna erabiltzea pentsatu genuen. Horri maiztasun normala esaten zaio (*normal frequency*), eta hori erabilita, emaitzak nabarmen hobetzen dira (% 10 inguru).

	Doitasuna	Estaldura	F-neurria
MI	36,84	75,91	49,61
MI^3	36,84	75,91	49,61
LR	36,45	71,54	48,30
Ji karratua	40,53	56,68	47,27
t neurria	39,92	66,93	50,01
Dice	42,36	70,35	52,88

12. taula. Bigramen emaitza hobetuak
(hitzen maiztasun 'normalak' erabiliz)

Dena den, doitasuna % 10 inguru igotzeko, estaldura beste hainbeste edo gehiago jaitsi behar da. Neurrien artean ez da alde nabarmenegirik sumatzen (Dice-ren koefizienteak ematen du F neurri onena).

3.2.1.2. Termino luzeagoen tratamendua ($2 < n < 5$)

Hiru edo lau hitzeko termino-hautagaiak ordenatzeko, bi estrategia erabili ditugu. Batean, sintagmaren bi osagai nagusien arteko elkartze-mailaren arabera ordenatu dira; bestean, berriz, osagai guztien elkartze-mailaren arabera. Lehen estrategian, termino habiatuen tratamendurako erabiltzen den gramatika erabili dugu, hau da, sintagmaren burua eta modifikatzailea identifikatzen dituena. Bigarren estrategian, osagai guztien elkartze-maila neurtzeko gai den ME neurria erabili dugu. Lehen bidetik, bigramen tratamenduan bezala, emaitzak hobetzen dira osagaien maiztasun normalak hartzen direnean. Lehenengo sei neurrien emaitzak (13. taula) lehen estrategiari jarraituz kalkulatu dira. Azken lerrokoak, aldiz, bigarren estrategiaren arabera. Emaitzarik onenak lehen estrategiaren bidez lortzen dira; Dice-ren koefizientearen bidez zehazki.

	Doitasuna	**Estaldura**	**F-neurrria**
MI	28,90	51,26	36,96
MI^3	28,90	51,26	36,96
LR	29,71	54,09	38,35
Ji karratua	28,90	51,26	36,96
t neurria	29,36	53,46	37,90
Dice	28,92	59,75	38,97
ME	19,24	70,13	30,20

13. taula. *n*-gramen emaitzak ($2 < n < 5$)

3.2.2. Termino bakunak

Hitz bakarreko terminoen *termhood*-a estimatzeko, hainbat neurri proposatu dira. IRn ohikoa den *tf-idf* da horietako bat. Dokumentumultzo bat izanik, neurri horrek haztatzen du terminoa zenbat aldiz agertzen den eta agertze hori dokumentuetan barrena nola banatuta dagoen; hau da, hitz bat dokumentu gutxitan baina horietan maiz agertuz gero, probabilitate handiagoa du termino izateko. Matsuo eta Ishizuka-k (2002) agerkidetzaren neurriak erabiltzen dituzte.

tf-idf modu egokian kalkulatzeko informazio nahikorik ez dugunez, Damerau-k (1993) proposatutako maiztasun erlatiboaren erlazioa (*Relative Frequency Ratio* - RFR) erabili dugu. Lortutako emaitzak 14. taulan azaltzen dira. Neurriak hitz anitzeko terminoenak baino hobeak izan arren, arrazoia ez da metodoa sofistikatuagoa dela, *Hapax Legomena* arazoa askoz txikiagoa dela baizik.

	Doitasuna	**Estaldura**	**F-neurria**
RFR	43,55	76,91	55,61

14. taula: Termino bakunen emaitzak

F neurria % 5 inguru hobetu da; doitasuna % 8 igo da, eta estaldura % 10 jaitsi.

Termino bakunen zein hitz anitzekoen emaitzak aztertuta, esan dezakegu, oro har, metodo estatistikoekin lortutako emaitzak ez direla oraindik sistema automatiko batek behar lukeen mailakoak. Nolanahi ere, eta etorkizuneko hobekuntzen zain, uste dugu lortu ditugun emaitzak onargarriak eta baliagarriak direla aplikazio erdiautomatikoetarako.

3.3. Aplikazioa

Helburu nagusia terminologia modu erdiautomatikoan erauztea izanik, berariazko aplikazio informatikoa eraiki da.

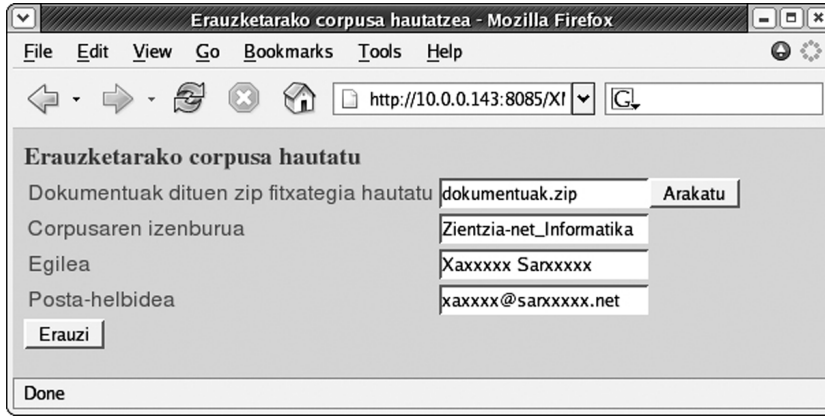
Alde metodologikoari garrantzi handia eman zaio, eta softwarearen ingeniaritzan puri-purian dauden bi paradigmatan oinarritu da diseinua:

- Hiru mailatako arkitektura logikoan oinarritzen da (*three-tier logical architecture*): erabiltzailearen interfazea, prozesatze logikoa eta datu-kudeaketa bereizi egiten dira
- Mailak banatzeko bezero-zerbitzari eredua hautatu da, *bezero arin* motakoa zehazki. Bezeroa erabiltzaile-interfazea kudeatzen duen web nabigatzailea da. Aplikazioaren funtsa, aldiz, web zerbitzari batek kudeatzen du (*Apache+mod_perl*). Azkenik, datu-kudeaketa XML datu-base zerbitzari baten bidez egiten da (Berkeley DB-XML)

Esan bezala, hiru osagai nagusi daude: corpusaren eraikitzailea, terminoen etiketatzailea eta emaitzak aztertzeko zein balioesteko nabigatzailea.

Corpusaren eraikitzailearen bidez, hainbat formatutako testuekin (*txt*, HTML, ...) corpus bat osa daiteke, corpus hori termino-erauz-

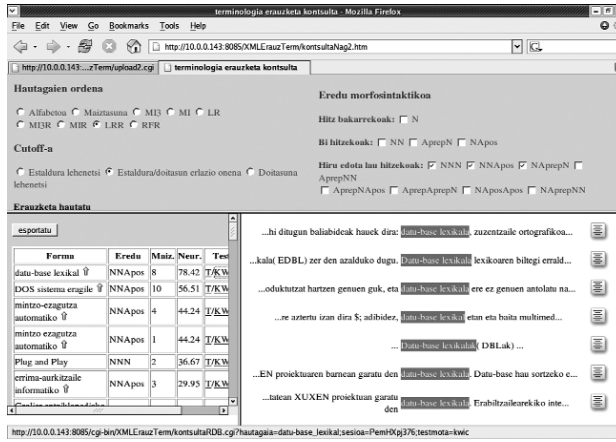
learen lehengaitzat erabil dadin. Testuak XML etiketak erabiltzen dituen tarteko formatu batera aldatzen dira, nazioarteko proposamenei jarraiki (TEI P4). Tarteko formatu hori ere bada beste moduluekiko integrazioa bideratzeko funtsa. 1. irudian, modulu horren interfazearen adibide bat dago.



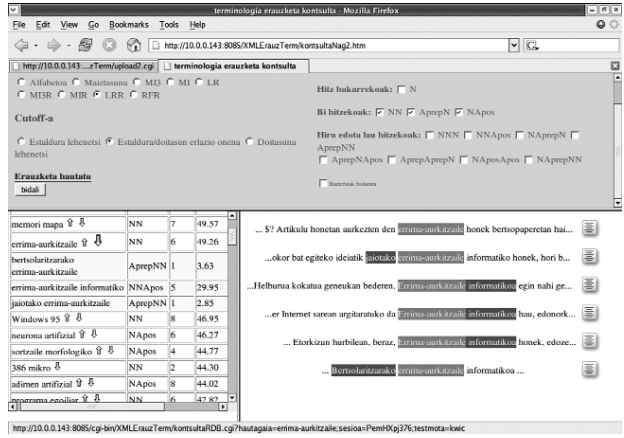
2. irudia. Corpusaren eraikitzailea

Bigarren moduluak, terminoen etiketatzaileak, azaldu ditugun teknika linguistikoei eta estatistikoei dagozkien prozesuak egiten ditu, eta, corpus bat aukeratuta, termino-hautagaiak bistaratzen ditu, pantailan aukeratutako neurri estatistikoaren bidez ordenatuta. Barnean hautatutako termino horiek eta dagozkien neurriak XML etiketen bidez islatzen dira.

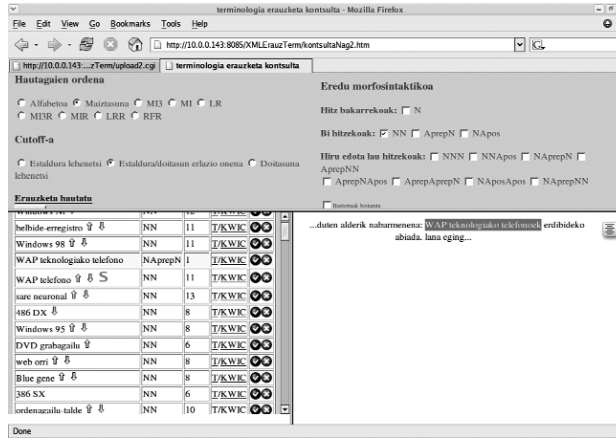
Hirugarren osagaia nabigatzailea da, eta, horren bidez, nagusiki, termino-hautagaien testuinguruak kontsulta daitezke. Horrez gain, eta emaitzen azterketa errazteko asmoz, termino habiatuak dagokien hautagai luzeenarekin erlazionaturik bistara daitezke, eta orobat zenbait aldaeraren forma kanonikoez. 2. irudian, azken bi moduluak bateratzen dituen interfaze nagusiaren adibide bat dago.



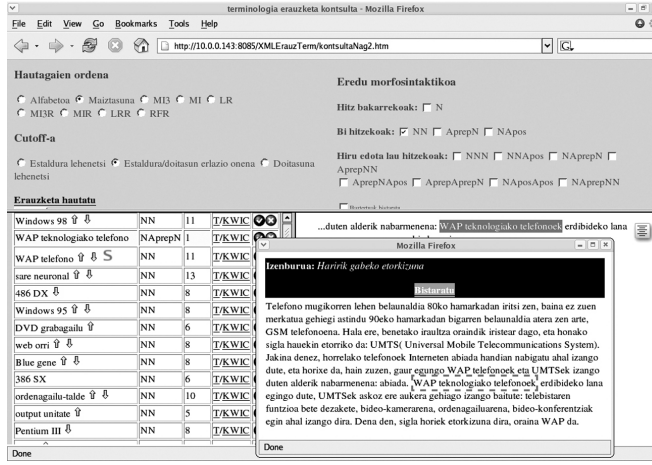
3. irudia. *Datu-base lexikal* terminoaren KWIC ikuspegia. Nabigatzailearen goiko atalean, emaitzak bistaratzeko aukerak daude: neurri estatistikoak, estaldura edo doitasuna lehenesteko aukerak eta eredu morfosintaktikoa



4. irudia. *Errima-aurkitzaile* terminoaren agerraldiak. Habiatua denean, dagokion termino luzeagoa ere nabarmendurik agertzen da. Eskuineko zerrendan, termino baten ondoko beheranzko gezia sakatuta, dagozkion termino-hautagai luzeenak bistaratzen dira (hau da, delako terminoa zein terminotan dagoen habiatuta)



5. irudia. *WAP telefono* ↔ *WAP teknologiako telefono* txertatze-aldaera nabigatzailean bistaratzeko modua. 'S' ikurrak adierazten du *WAP telefono*-ren aldaera sintaktiko bat detektatu dela



6. irudia. *WAP teknologiako telefono* terminoaren agerraldia 'Paragrafoa' ikuspegian (konkordantzia-lerroaren eskuinaldeko ikonoa sakatu behar da paragrafoa bistaratzeko, edo termino-zerrendako 'T' esteka). Dokumentu osoa ikusteko, paragrafoaren leihoan ageri den 'Bistaratu' esteka sakatu behar da

4. Ondorioak eta etorkizuneko lanak

Euskarazko testuetatik abiatuta terminologia erauzten laguntzeko tresna aurreratu bat aurkeztu dugu. Hizkuntz ingeniaritza, metodo estatistikoak eta teknologia informatiko modernoa integratzen ditu aplikazio honek, eta terminologia-lanetan ari direnei orain arte izan ez duten laguntza eskaini nahi die.

Etorkizunari begira, hainbat hobekuntza eta asmo berri dugu esku artean:

- Oraingo sistema hobetzea: etiketatzaileak dituen mugak gainditzea, postposizioen lanketa integratzea, neurri estatistikoak fintzea eta ebaluazio-corpus osoago bat biltzea. Horiez gain, erabilera errealean aurkitzen diren arazoak eta eragozpenak zuzentzea, eta zein hobekuntza egin litekeen aztertzea
- Semantika barneratzea, Euskal WordNet ezagutza-base semantikoa garatu ahala. Semantikaren integrazioaren bidez, normalizazio semantikoari aurre egiteko aukera izango dugu, baita emaitzak semantikoki multzokatuta aurkezteko ere
- Terminoen arteko erlazioetan oinarrituta (aldaerak, termino habiatuak...), kontzeptuerlazioak erauztea
- Erauzle elebiduna garatzea. Testu paraleloetatik bi hizkuntzatako terminoen arteko baliokidetzak, hau da, termino parekatuak lortzea da tresna horren helburua. Dagoeneko zenbait lan egin ditugu alor honetan, eta garapen hori urrats garrantzitsua izango da itzulpenari zein normalizazioari begira

5. Esker ona

Lan hau Hizking21 proiektuaren baitan egin da (www.hizking21.org), eta Eusko Jaurlaritzaren *Etortek* eta *Saiotek* programen diru-laguntza jaso du.

6. Bibliografia

ADURIZ, I., ALEGRIA, I., ARTOLA, X., EZEIZA, N., SARASOLA, K. & URKIA, M. 1997. "A spelling corrector for Basque based on morphology". In *Literary & linguistic computing*, vol. 12, no. 1. Oxford: Oxford University Press. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1000911636/publikoak/97LITER_Z.ps

ALEGRIA, I., ARTOLA, X. & SARASOLA, K. 2001. "Hizkuntzaren tratamendu automatikoa: aplikazioak, tresnak, baliabideak eta oinarriak". In: *Euskonews*. <http://suse00.su.ehu.es/euskonews/0110zbk/frgaia.htm>

BEESLEY, K.R. & KARTTUNEN, L. 2003. *Finite State Morphology*. CSLI. Standford University.

BLAHETA, D. & JOHNSON, M. 2001. "Unsupervised learning of multi-word verbs." In *Proceedings of the 39th Annual Meeting of the ACL*. 54-60. Tolosa.

BOURIGAULT, D. 1994. *LEXTER, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances a partir de textes*. Ph.D. Thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

BOURIGAULT, D. 1996. "Lexter, a Natural Language Processing Tool for Terminology Extraction." In *Proceedings of 7th EURALEX International Congress*.

BOURIGAULT, D. & JACQUEMIN, C. 1999. "TERM EXTRACTION+TERM CLUSTERING: An Intergrated Platform for Computer-Aided Terminology." In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Bergen, Norvegia.

CABRÉ, T. 2001. "Consecuencias metodológicas de la propuesta teórica (I)." In *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica*. Bartzelona: IULA-UPF. 27-36.

CHURCH, K.W. & HANKS, P.P. 1989. "Word association norms, mutual information and lexicography." In *Proceedings of the 27th Annual Meeting of the ACL* (pp. 76-83). Vancouver.

DAILLE, B. 1995. "Combined approach for terminology extraction: lexical statistics and linguistic filtering". In *UCREL Technical Papers*, 5, University of Lancaster.

DAILLE, B., HABERT, B., JACQUEMIN, C. & ROYAUTÉ, J. (2000) "Empirical Observation of Terms Variations and Principles for their Description." In *Terminology*, 3(2), 197-258. Amsterdan: John Benjamins

DAMERAU, F.J. 1993. "Generating and evaluating domain-oriented multi-word terms from texts." In *Information Processing & Management*, 29, 433–447. Elsevier

DIAS, G., GUILLORÉ, S., BASSANO, J.C. & LOPES, J.G.P. 2000. "Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?" In *Proceedings of Recherche d'Informations Assistée par Ordinateur* 1-20. Paris.

DIAS, G., GUILLORÉ, S. & LOPES, J.G.P. 1999. "Mutual Expectation: a Measure for Multiword Lexical Unit Extraction." In *Proceedings of VExTAL Venezia per il Trattamento Automatico delle Lingue*. Universitá Cá Foscari. Venezia, Italia.

DUNNING, T. 1994 "Accurate Methods for the Statistics of Surprise and Coincidence." In *Computational Linguistics* 19(1): 61-74. Cambrigde, Mass.: The MIT Press.

ELOSEGI, A. 2002. *Zuzenbideko euskal hizkera berezia. Lege-corpus bateko terminologiaren azterketa linguistikoa eta terminologikoa*. Euskal Herriko Unibertsitatea. Euskal Filologia Saila.

ENSUNZA, M., ETXEBARRIA, J.R. & ITURBE, J. 2002. *Zientzia eta teknikarako euskara. Zenbait hizkuntza-baliabide*. Bilbo: Udako Euskal Unibertsitatea

ESTOPA, R. 2001. "Elementos lingüísticos de las unidades terminológicas para su extracción automática" In: CABRÉ, M.T et al. (ed.): *La Terminología científico-técnica: reconocimiento, análisis y extracción formal y semántica*. 67-80. orr. Bartzelona: IULA-Institut Universitari de Lingüística Aplicada- Universitat Pompeu Fabra.

EVERT, S. 2001. "On lexical association measures." <http://www.collocations.de/EK/am-html/>

EVERT, S. 2004. *Computational Approaches to Collocations*. <www.collocations.de>

EZEIZA N., ADURIZ I., ALEGRIA I., ARRIOLA J.M. & URIZAR R. 1998. "Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages." In *COLING-ACL'98,* Montreal. <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1000911659/publikoak/98ACLdesanb.ps>

IXA 2002. EusWordNet. <http://sisx03.si.ehu.es/tresnak/wei/wei_mysql_euskaraz.html>

JACQUEMIN, C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Mass.: The MIT Press.

JACQUEMIN, C. 1999. "Syntagmatic and Paradigmatic Representation of Term Variation" In *ACL'99*, p. 341–348.

JACQUEMIN, C. 1999. "NLP for Term variant extraction, synergy between morphology, lexicon and syntax." In T. Strzalkwski (ed.), *Natural Language Processing Information Retrieval*. 25-74. Kluwer, Boston, MA..

JUSTESON, J. 1993. "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text." In: *IBM Research Report, RC 18906 (82591)*.

KAGEURA, K. & UMINO, B. 1996. "Methods of Automatic Term Recognition." In *Terminology*. 3(2), 259-289. Amsterdan: John Benjamins.

LAPATA, M. & LASCARIDES, A. 2003 "Detecting Novel Compounds: The Role of Distributional Evidence". In *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics* (pp. 235-242). Budapest.

MANNING, Ch., & SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

MATSUO, Y. & ISHIZUKA, M. 2000. "Keyword extraction from a document using word co-occurrence statistical information". In *Transactions of the Japanese Society for Artificial Intelligence*, 17(3), 217-223.

MAYNARD, D. & ANANIADOU, S. 2000. "Trucks: A Model for Automatic Multi-Word Term Recognition." In: *Journal of Natural Language Processing*, 8(1), 101-126.

NAKAGAWA, H. & MORI, T. 2003. "Automatic Term Recognition based on Statistics of Compound Nouns and their Components". In *Terminology*, 9(2), 201-219. Amsterdan: John Benjamins.

PEREZ GAZTELU, E. 1995. *Koldo Mitxelena Elissalt. Egitasmoa eta egitatea*. Errenteria: Erenteriako Udala

SMADJA, F. 1993. "Retrieving Collocations from Text: XTRACT." In: *Computational Linguistics*, 19(1) 143-177. Cambridge, Mass.: The MIT Press.

URIZAR R., EZEIZA N. & ALEGRIA I. 2000. "Morphosyntactic structure of terms in Basque for automatic terminology extraction." In: *Proceedings of the 9th EURALEX International Congress*. 373-382. Stuttgart. <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1012399093/publikoak/euralex2000.pdf>

VIVALDI, J. 2001. "Elaboración de una aplicación automática de reconocimiento y extracción de información terminológica en textos

de dominios restringidos". In: CABRÉ, M.T et al. (ed.): *La Terminología científico-técnica: reconocimiento, análisis y extracción formal y semántica*. 67-80. orr. Bartzelona: IULA-Institut Universitari de Lingüística Aplicada- Universitat Pompeu Fabra.