INTERNET, CORPUSAK ETA TERMINOLOGIA: INTERNETETIK ESPEZIALITATE-CORPUSAK ERAUZTEKO TEKNIKAK ETA HORIEN EBALUAZIOA

Antton Gurrutxaga, Igor Leturia, Eli Pociello, Iñaki San Vicente, Xabier Saralegi
(Elhuyar Fundazioa)

AURKIBIDEA

1. Sarrera

Datuetan oinarritutako hiztegigintza eta, zehazki, terminologia-lana hasia da gurean prozedura estandarra izateko bidea egiten. Zorionez, esan behar genuke. Duela 30 urte euskara espezialitate-diskurtsoan erabiltzen hasi ginenetik sortu den erabilera erreala alferrik izan ez dela pentsatuta, argi dago espezialitate-arloetako testu-produkzioa behatzea eta hortik terminologia-lanerako datuak ateratzea merezi duela, edo, are gehiago, horri ekitea ezinbestekoa dela, baldin heldutzat jo daitekeen terminologia-lana egingo badugu.

Gaur egun, datuetan oinarritutako terminologia-lan estandarrean, terminoak "testuetan" daudela onartzen da, hau da, hizkuntza-jardueraren emaitza diren produkzio idatzi zein ahozkoetan (Cabré 2001: 17).

Beraz, espezialitate-corpusak behar ditugu; bestetik, terminoen erauzte-lana automatikoki egiteko tresnak ere bai. Euskara ari da bere burua prestatzen eta janzten datuetan oinarritutako terminologia-lana egiteko. Esaterako, EHUko IXA taldeak eta Elhuyar Fundazioko I+G unitateak *Zientzia eta Teknologiaren Corpusa*[1] (Areta *et al.* 2007) eta *Erauzterm* termino-erauzle automatikoa (Alegria *et al.* 2004) garatu dituzte hamarkada honetan. Dena den, iritsi da garaia, gure ustez, azken hamarkadan corpusgintzan aukera berriak eta aldaketa nabariak ekarri dituen errealitate bat kontuan hartzen hasteko: Internet.

Lan honen helburua bikoitza da. Batetik, Internetek corpusgintzan ekarri duen aldaketa aztertu nahi dugu. Baina gogoeta hutsetik haraindi ere joan nahi genuke, corpusgintza esperimentala egin, eta ahal den objektibokien ebaluatu. Horretarako, lehenik, Internetetik espezialitate-corpus batzuk eratu ditugu Elhuyar I+Gk garatu duen Co3 tresnaren bidez. Gero, corpusak espezialitate-arloaren aldetik karakterizatzeko, bertatik automatikoki erauzitako terminoak ebaluatuko ditugu, nagusiki hiztegien bidez.

2. Internet eta corpusak

XX. mendean corpusek hizkuntzalaritzan, hiztegigintzan eta hizkuntzaren prozesamenduaren teknologietan ekarri duten aldaketa nabaria da. Corpusen egiteko behinena da hizkuntzaren erabilera errealaren ebidentziak biltzea, eta, beraz, corpus ideal bat litzateke aztertu nahi den hizkuntza-erabileraren adierazgarria dena (Biber 1993: 243). Horrexegatik eman zaio garrantzi handia corpus-diseinuari, horrek berma lezakeelakoan corpusaren adierazgarritasuna, edo, behintzat, horretara albait hurbiltzeko aukera eman.

Corpus-diseinu jakin bat gauzatzeak kostu eta zailtasun handiak ekarri ohi ditu berekin, batez ere oreka eta tamaina handia nahi direnean. Seguru aski, hori izan da corpusgintza

[1] http://www.ztcorpusa.net [2010-03-29]

"oportunista" sortu izanaren arrazoia; arreta handiagoz begiratzen zaio corpusean sar litezkeen obren eskuragarritasunari eta horiek prozesatzeko erraztasunari.

2.1. Zergatik Internet corpusgintzan?

Bi arrazoi nagusi daude Internetek corpusgintzan piztu duen interesa eta izan duen eragina ulertzeko.

Lehena zeharka azaldu berria dugu: praktikotasuna. Interneten testu-kantitate handia dago, digitalizatuta eta eskura; gainera, Interneten ere argitaratzen diren testuak gero eta gehiago dira. Ezin ukatuzkoa da corpuserako testuak Interneten egoteak corpusgintzari mesede egiten diola, eta bidea errazten. Esan daiteke interesgarria dela webetik corpusak automatikoki eratzeko tresnak garatzea.

Baina, horrez gain, badira arrazoi linguistikoak ere corpusgintzatik Internetera begiratzeko: Internet ezin ukatuzko errealitate "linguistikoa" ere bada. Interneten bakarrik argitaratzen diren testuak gero eta ugariagoak dira, eta ezaugarri bereziak dituzte. Horietako asko, gainera, bat-bateko diskurtsoak dira, eta hizkuntzaren erabilera erreala islatzen dute (foroak, blogak...). Esan daiteke interesgarria dela webaren alderdi linguistiko bereziak aztertzea.

2.2. Interneten erabilera corpusgintzan

Bi eratako ikuspegi daude Internet corpus-lanean erabiltzeko. Lehena, Internet zuzenean corpus bat balitz bezala kontsultatzea (*web as corpus*); bigarrena, Internet corpusa eratzeko edo elikatzeko testu-iturritzat erabiltzea (*web for corpus*)[2].

2.2.1. Internet corpustzat (*web as corpus*)

Jakina da web-bilatzaileak bere horretan ere erabili ohi direla hitzen agerpenen informazioa eskuratzeko. Baina kontuan izan behar dugu tresna horien helburua ez dela berez horrelako datuak ematea, dokumentuak bilatzea baizik. Ildo horretatik uler dezakegu A. Kilgarriffek, *googleology* izendatu duen praktikaz mintzatu zenean, "bad science" dela esan izana (Kilgarriff 2007).

Horregatik, hainbat tresna egin dira Interneten hitz bat edo batzuk bilatzeko aukera ematen digutenak, baina emaitza, dokumentu-zerrendara mugatu beharrean, dokumentu horietako agerpenak euren testuinguruetan erakusten dituztenak. Adibide batzuk: *WebCorp*[3] (Kehoe & Renouf, 2002), *WebCONC*[4] edo *KWiCFinder*[5] (Fletcher 2001).

Tresna horiek, ordea, ez dute euskararekin lan egiten, eta horregatik garatu du Elhuyar Fundazioak *CorpEus* proiektua[6], Internet euskarazko corpus erraldoi gisa baliatzea helburu duena (Leturia *et al.* 2007). Bilaketa euskarazko dokumentuetan eta euskararen berezitasunetarako garatuta egiten du (esaterako, lema baten forma flexionatuak aurkitzen ditu).

---

[2] Egia da gaur egun *web as corpus* izendapenaren barnean sartu ohi dela Internet corpusgintzan era batera edo bestera erabiltzeko ekimena, baina bereizketa egiteak lagundu egingo digu kontzeptuak azaltzen, eta hemen behintzat erabili egingo dugu.

[3] http://www.webcorp.org.uk/ [2010-03-29]

[4] http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?sprache=en [2010-03-29]

[5] http://www.kwicfinder.com/KWiCFinder.html [2010-03-29]

[6] http://www.corpeus.org [2010-03-29]

Egia da bide horrek badituela koska batzuk. Esaterako, bilatzaileek eskaintzen dituzten aukeren eta ezartzen dituzten rankingen mendekoa da hein batean. Bestetik, corpusa ez dago sailkatuta eta linguistikoki etiketatua, eta horrek muga batzuk ezartzen dizkio bilaketari. Baina abantailak ez dira nolanahikoak: uneko datu eguneratuak eta kostu txikia (Renouf 2007: 42).

2.2.2. Internet corpusgintzarako (*web for corpus*)

Internetez baliatzeko bigarren bidea *off-line* corpusak eratzea da. Hori bi modutara egin ohi da: *crawling* bidez, eta hitz batzuetatik abiatuta bilatzaileak erabiliz. Azkenaldian, bigarren bidea nagusitu dela dirudi, ziurrenik metodologia hori erabiltzen duen *BootCaT* tresna (Baroni & Bernardini 2004) lan mota honetarako ia *de facto*-ko estandarra bihurtu delako. Tresna horren bidez eratu dira, adibidez, *Wacky* proiektuaren barruko *ukWac*, *itWaC* eta *deWaC* corpusak (2 mila milioi hitz ingurukoak)[7]. Gainera, *Corpus building for minority languages* gunean[8], K. P. Scannell-ek *An Crúbadán web crawler*-aren bidez osatutako 419 hizkuntzaren corpusen berri ematen du, eta, horien artean, euskarazko corpusen datu batzuk ematen ditu (Scannell 2007). Azkenik, dagoeneko 5,5 mila milioi hitzetara heldu den *BiWec* proiektua da aipatzekoa (Pomikálek *et al.* 2009).

Beste batzuetan gertatzen den bezala, tresna horiek ez dute euskararen ezaugarrietarako balio, eta Elhuyar Fundazioak Co3 proiektua abiarazi du (*Comparable Corpus Collector*), hurrengo atal batean xeheago aurkeztuko duguna.

2.3. Zenbait kezka

Ikusi dugunez, webaren sorrerak aukera berriak eskaini dizkio corpusgintzari eta corpus-hizkuntzalaritzari, baina orain arte ez bezalako erantzunak eskatzen dituzten galdera berriak eginarazi ere bai. Aukerak azalduak ditugu. Baina, hizkuntzaren azterketaren ikuspegitik, zenbait galdera eta eztabaidagai jarri ditu aukera horrek mahai gainean.

2.3.1. Adierazgarritasunaren arazoa

Adierazgarria ez izatea da Internet corpustzat hartzeari jarri izan zaion eragozpenik handiena. Corpus kontzeptuaren definizio zehatz batean oinarrituta dago hori: "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, J. 2005). Interneten dagoen populazioa ezezaguna da, eta une oro aldatzen ari da, gainera. Internet bera, hortaz, ezin da corpusa denik esan (Sinclair, J. 2005).

Internet corpustzat aldarrikatzen dutenak definizio zabalago batetik abiatzen dira: "A corpus is a collection of texts when considered as an object of language or literary study" (Kilgarriff *et al.* 2003). Onartzen dute adierazgarritasunaren arazoa, baina haien ustez galdera litzateke ea Internet corpus diseinatuak baino "ez-adierazgarriagoa" den, zeren corpus horietan ere populazioa ezin baita erabat zehaztu. Alde horretatik, beste corpusak bezain ez-adierazgarria litzateke Internet.

Nolanahi ere, Internet "corpuserako" delako ikuspegitik, galdera da Interneteko dokumentuak erabiliz corpus adierazgarririk edo orekaturik behintzat egin daitekeen. Bi alderdi hartu behar dira kontuan: Interneten corpus orekatu bat eratzeko adinako produkzio-

[7] http://wacky.sslmit.unibo.it/doku.php [2010-03-29]
[8] http://borel.slu.edu/crubadan/stadas.html [2010-03-29]

dibertsitaterik badugun; eta nola sailkatuko ditugun dokumentuak, corpus-diseinu baten araberako proportzioak lortzeko.

2.3.2. Datuen "ezegonkortasuna" edo "errepikaezintasuna"

Internet une oro ari da aldatzen. Hortaz, une jakin batean lortzen ditugun datu linguistikoak "iraunkorrak" ez izateko arriskua aipatu izan da, eta horrek analisi linguistikoan dakarren desabantaila, ateratzen ditugun ondorioak egonkorrak ez izatea, alegia (Renouf 2007: 42). Gainera, hizkuntza-teknologietan egiten den ikerkuntzan, esperimentuen "berregingarritasuna" bermatu behar da, eta horrek eskatzen du corpus egonkor bat edo modu kontrolatuan hazten den bat erabiltzea (Lüdeling *et al.* 2007). Alde horretatik, autore horien ustez ez da oso egokia weba ebaluazio-corpustzat erabiltzea, edo gainditu gabeko arazoak sortzen ditu oraindik.

2.3.3. Hizkuntzaren "kalitatearen" auzia

Testu inprimatuetan ez bezala, Interneten orraztu gabeko testu asko dago, erregistro informalean idatziak eta abar. Horrek kezka sortu ohi du Internet hizkuntza aztertzeko datu-iturritzat erabiltzea proposatzen denean, eta askotariko iritziak entzuten dira. Oro har, eta sinplifikatuta, esan daiteke bi ikuspegi daudela Interneteko hizkuntza-kalitatearen kontua epaitzerakoan:

- Interneteko orri askotako hizkuntza zaindu gabea da, kalitate eskasekoa → ez du merezi kontuan hartzea; kaltegarria eta arriskutsua da corpuseratzea
- Interneten egiten diren komunikazio-ekintzetan erabiltzen den hizkuntza euskara "erreala" ere bada → hizkuntzalaritzaren aztergaia da, eta datutzat har daitezke (edo hartu behar dira)

Corpus-hizkuntzalaritzaren ikuspegitik, kalitatea ez litzateke testuak corpuseratzeko erabili beharreko iragazki bat; helburua hizkuntza aztertzea bada, zaila da defendatzea testu batzuk *a priori* aztergaitik kanpo utzi behar direla. Corpusaren helburua estandarrerako ereduak lantzea denean ere, corpusa *a priori* norbaiten kalitate-irizpide batzuen arabera "murrizten" badugu, aurrez hartzen dugu baliabidea moldatzeko erabaki bat, gero baliabide hori ustiatu eta estandarrerako proposamenak lantzeko. Proposamenak aurrez baldintzatuta leudeke; horretarako, ez da corpusik behar.

Beraz, uste dugu kalitatearen arazoa ez dela Internetek corpusgintzan izan dezakeen erabilera zalantzan jartzeko adinako arrazoia. Laburbilduz, datuak aztertzea ez da datuak eredutzat hartzea.

3. Internetetik espezialitate-corpusak eratzeko teknikak

Gure iritziz, garaia da Internetek corpusgintzarako eta hiztegigintzarako sortu dituen espektatibak egiaztatzen hasteko. Espezialitate-arloan ari garela, esaterako, galde liteke ea arlo bateko hiztegi terminologiko bat eratzeko liburuak eta dokumentuak (bakarrik) hustu beharrean, Interneten (ere) oinarritu gaitezkeen. Aurretik, ordea, Internetik nahi ditugun ezaugarriak dituzten corpusak eratzeko tresnak garatu behar ditugu.

Izan ere, Internet corpusgintzarako testu-iturri sistematiko gisa erabiliko badugu, beharrezkoa da dokumentuak karakterizatzea eta sailkatzea, dela corpus orokor orekatu bat dela corpus berezi edo espezialitateko bat eratzeko. Corpus-diseinuan, arloa/eremua/gaia edo

erregistroa/generoa eta antzeko parametroak erabili ohi dira dokumentuak modu kontrolatu eta orekatuan hautatzeko.

Artikulu honetan, arloaren parametroa landuko dugu. Internetik espezialitate-corpusak automatikoki eratzeko Elhuyarrek garatu duen tresna aurkeztuko dugu, eta tresna horrekin egindako esperimentu batzuk azalduko ere.

Ikertze-ildo honetan, Elhuyarrek ez du euskarazko corpusgintza bakarrik kontuan izan. Bada, hainbat arrazoi daude corpus konparagarrien arloan barneratzeko. Euskarazko corpus paraleloak urriak dira iturburu-hizkuntza gaztelania ez bada. Beraz, aukera bat izan daiteke corpus konparagarriak ustiatzea, hau da, itzulpenak ez diren baina ezaugarri batzuk (arloa, generoa, data...) partekatzen dituzten testuez osatutako corpus eleaniztunak ustiatzea. Horretarako garatzen ari gara *Comparable Corpus Collector* - *Co3* tresna eta, horrelako corpusak terminologiarako ustiatzeko, *AzerHitz* erauzlea (Saralegi *et al.* 2008).

3.1. *Comparable Corpus Collector* - Co3

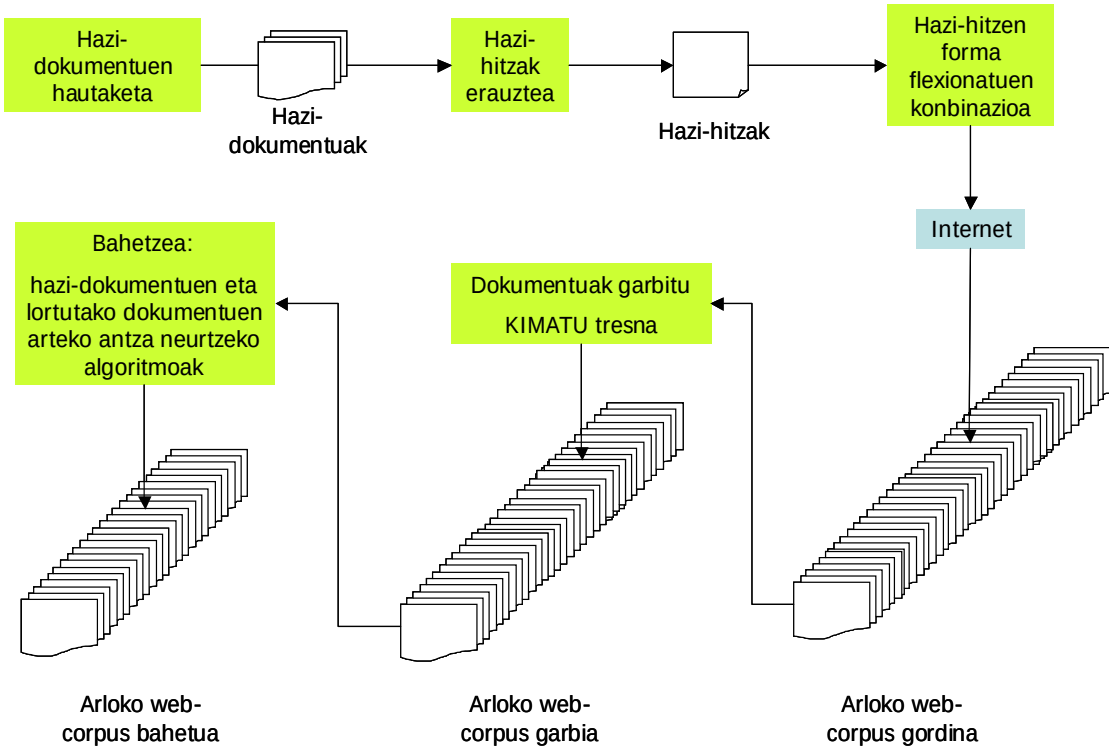
Co3 Internetik espezialitate-corpus konparagarriak eratzeko tresna automatikoa da (Leturia *et al.* 2009). Horrelako tresna batek gai izan behar du, lehenik, webeko dokumentuen hizkuntza ezagutzeko, eta arlo jakin bateko testuak identifikatzeko. Beste egiteko batzuk dokumentuak iragaztearekin, formatua bihurtzearekin eta garbitzearekin erlazionatuta daude. Estrategia hau erabiltzen du *Co3*k nahi dugun arloko testuak identifikatzeko:

- "Hazi-hitzen" (*seed words*) eta bilatzaileen bidezko kontsulten sistema estandarrean oinarritzea (Baroni & Bernardini, 2004)
- "Hazi-dokumentuetatik" (*seed documents*) ateratzea hazi-hitzak, automatikoki

Dokumentuak garbitzeko, *Kimatu* tresna erabiltzen dugu (Saralegi & Leturia 2007). Amaieran, azken bahetze bat egiten da, hazi-dokumentuen eta lortutako dokumentuen arteko antza neurtzeko algoritmoak erabiliz (Saralegi & Alegria 2007). 1. irudian ageri da prozesuaren diagrama.

4. Lan esperimentala

Esan bezala, esperimentuen helburua da Internet espezialitate-lexikografiarako datu-iturri erabilgarria izan daitekeen aztertzea. Horretarako, zenbait espezialitate-arlotako euskarazko web-corpusak eratuko ditugu, eta horiek espezialitate-arloaren aldetik duten "kalitatea" ebaluatuko dugu; horretarako, corpusak *Erauzterm* tresnaren bidez prozesatuko ditugu, eta automatikoki erauzitako termino hautagaiak hiztegien bidez ebaluatuko.



1. irudia. Co3ren corpusgintza-prozesua.

4.1. Web-corpusak eratzea

*Co3*ren bidez, hiru arlo hauetako euskarazko web-corpusak eratu ditugu: atomo- eta partikula-fisika, bioteknologia eta informatika. Arlo bakoitzean erabilitako hazi-hitzen kopuruak, hurrenez hurren, 63, 60 eta 105 dira.

Zientzia zein teknologiako arloak aukeratu ditugu, "hedadura" desberdinekoak, a priori behintzat, tamaina desberdineko corpusak sortuko lituzketenak. Corpusak eratzean ez dugu hitz-kopurua aurrez mugatu. Corpusei hazten utzi diegu, hazte-abiadura ia hutseratu arte.

4.2. Termino-erauzketa

Hurrengo urratsean, *Erauzterm*en bidez corpusak prozesatu eta terminoak automatikoki erauzi ditugu. *Erauzterm*ek zenbait neurri estatistiko erabil ditzake erauzitako termino hautagaien rankingak eratzeko. Azterlan honetan, LLR (*log-likelihood ratio*, edo egiantz-arrazoia) neurriaren araberako rankingak erabili ditugu, horrekin lortu baititugu doitasun eta estaldura onenak.

4.3. Terminoak baliozkotzea

Erauzketaren emaitza diren termino hautagaietatik, automatikoki baliozkotu dira Elhuyar Fundazioaren ZTH-*Zientzia eta Teknologiaren Hiztegi Entziklopedikoa*n daudenak (http://zthiztegia.elhuyar.org), eta hiztegian duten arlo-informazioa gorde. Orobat *Euskalterm*ekin, Interneteko kontsultagunea erabiliz (http://www1.euskadi.net/euskalterm/indice_i.htm), eta arlo-sailkapena ZTHrenarekin mapatuz. Azkenik, bi baliabide horietan ez dauden hautagai batzuk ere eskuz ebaluatu ditugu, eta termino berri batzuk baliozkotzat jo eta sailkatu.

4.4. Emaitzak eta analisia

Aurreko bi lan horien emaitza orokorrak 1. taulan ageri dira.

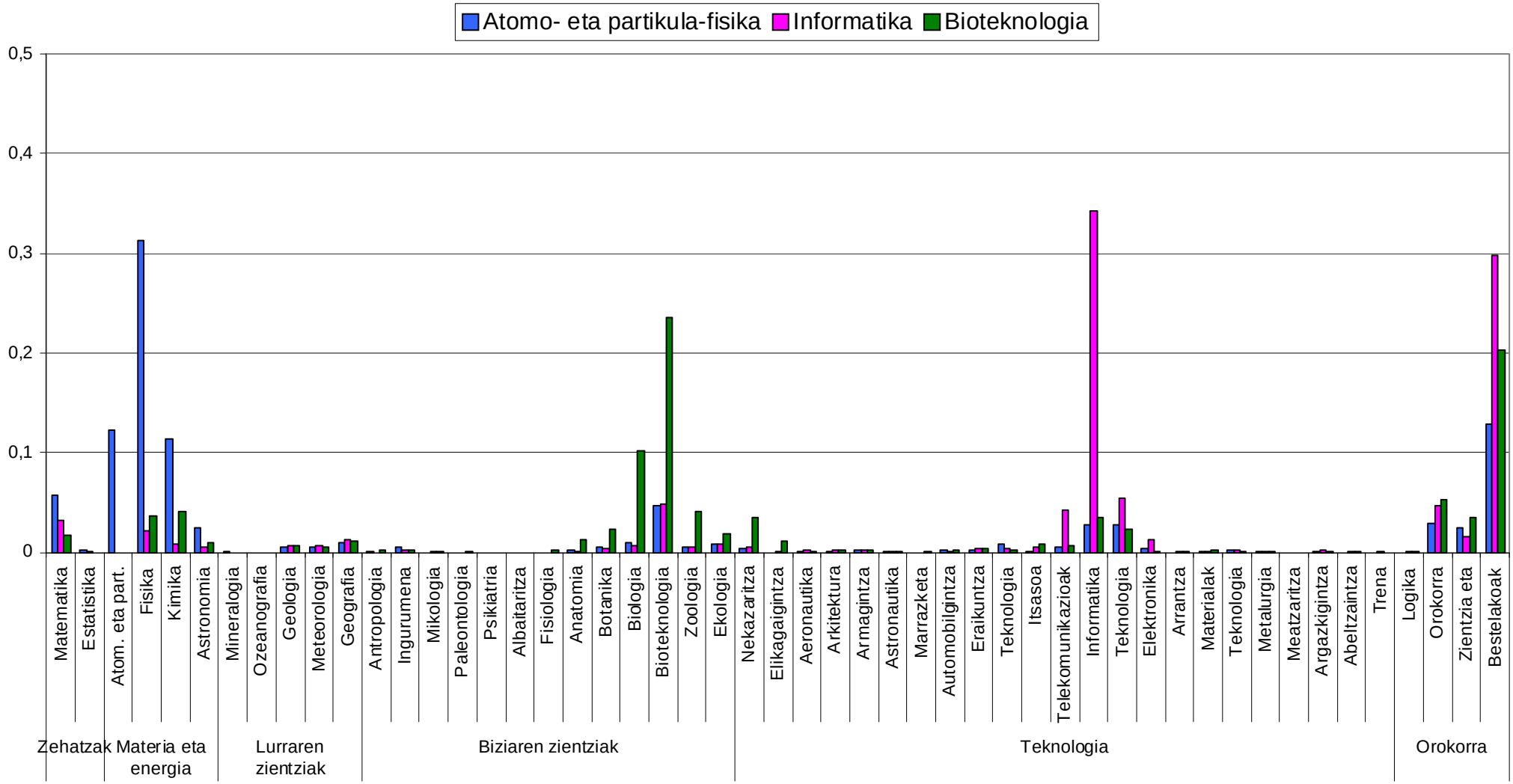
CORPUSA	ATOMO- ETA PARTIKULA-FISIKA	BIOTEKNOLOGIA	INFORMATIKA
Hazi-dokumentuen corpusa	32 dok. 26.164 hitz	55 dok. 41.496 hitz	33 dok. 34.266 hitz
Web-corpusaren tamaina	48 domeinu 310 orri 320.212 hitz	68 domeinu 358 orri 578.866 hitz	485 domeinu 1.810 orri 2.514.290 hitz
Erauzitako termino-kopurua	46.972	34.910	163.698
Hiztegien bidez baliozkotuak	6.432	6.524	8.137
Lehen 10.000 hautagaiak			
Hiztegien bidez baliozkotuak	2.827	2.403	2.755
Eskuz ebaluatuak	869	628	904
Positiboak	628	432	512
Negatiboak	241	196	392

1. taula: Co3-ren bidez eratutako espezialitate-corpusen ezaugarriak, eta termino-erauzketaren emaitzak.

Hurrengo ataletan, erauzitako lehen 10.000 terminoen zerrendak hartuko ditugu kontuan ebaluazioan.

4.4.1. Erauzitako terminoen arlokako banaketa

Web-corpusen espezializazio-profila zehazteko, bakoitzetik erauzitako baliozko terminoak zein arlotakoak diren eta nola banatuta dauden aztertu dugu lehenik. Hurrengo grafikoan bildu dugu azterlan horren emaitza. Lehen 10.000 hautagaien datuak dira, eta ordenatuen ardatzak doitasuna adierazten du.



2. irudia. Hiru web-corpusetatik erauzitako terminoen arlokako banaketa.

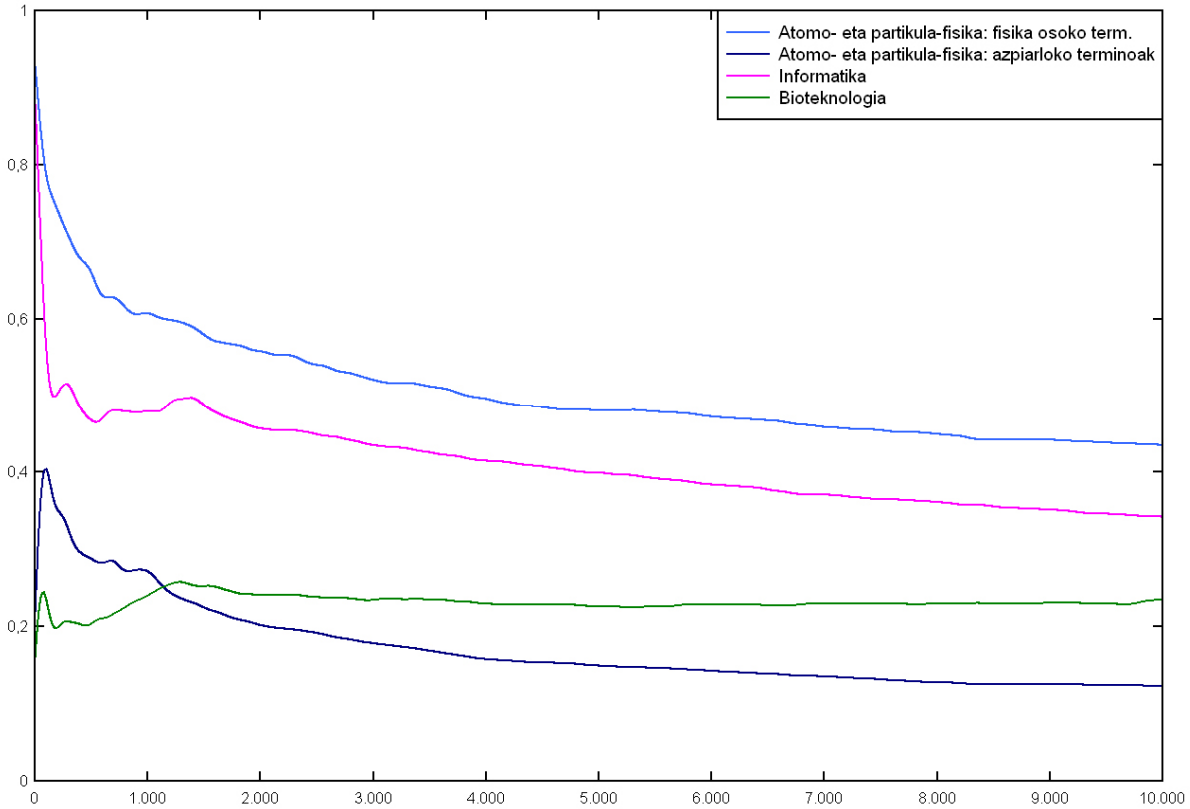
Oro har, bistan da hiru corpusetatik erauzitako terminoen banaketek dagokien arloetan edo oso gertukoetan dituztela puntako balioak, bakoitza espero zen arloko corpus espezializatua den seinale, neurri handian behintzat.

Fisikaren kasuan, bereizi egin ditugu atomo- eta partikula-fisikako terminoak eta fisikaren gainerako arlokoak. ZTH hiztegian, "Fisika" arlo-marka duten terminoetatik, % 24 dira atomo- eta partikula-fisikaren azpiarlokoak. Web-corpusaren erauzketan, % 39 dira azpiarlo horretakoak. Beraz, esan daiteke eratu dugun web-corpusa, fisika-arlokoa izateaz gain, azpiarlo aldetik ere espezializatua dela hein batean.

Bestetik, lehen datu batzuk ditugu pentsatzeko bioteknologiako web-corpusa dela arlo aldetik bereizi gabeena dena, terminoen banaketa lauagoa delako. Bestetik, corpus horren erauzketan ere, eta batez ere informatikarenean, nabarmentzekoak dira zientzia eta teknologiakoak ez diren arloetako terminoen ehunekoak. Bi emaitza horien esplikazioaren gakoa arloen izaeran egon daiteke. Teknologiak dira, hau da, aplikazio-arloekin erlazioa dute, eta ulertzekoa da testuetan agertzea arlo horietako edukiak eta, beraz, terminoak. Hala ere, uste dugu hazi-dokumentuak aukeratzeko prozesuak ere zerikusia izan dezakeela horretan, eta hori hobetzeko teknikak azterkizuntzat ditugu.

4.4.2. Doitasuna

Web-corpus bakoitzetik erauzitako terminoetatik dagokion arlokoak zenbat diren neurtuz, corpus bakoitzaren arlo-doitasuna ere hazta daiteke. Hurrengo irudiak erakusten digu termino-erauzketa bakoitzaren doitasuna nola aldatzen den erauzitako terminoen rankingetik kontuan hartzen den termino-kopurua handitu ahala:



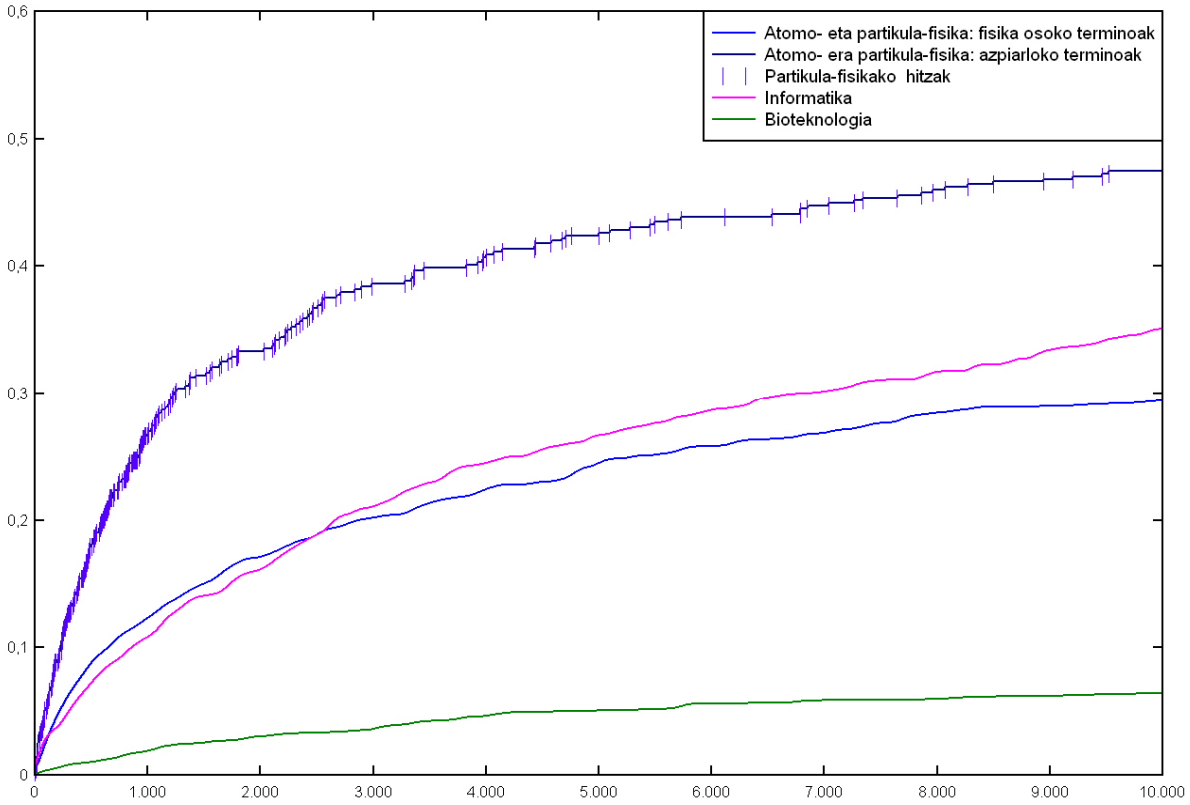
3. irudia. Termino-erauzketen arlo-doitasuna (baliozko termino guztiekiko).

Irudi horrek baieztatu egiten ditu aurrekoan sumatu ditugun alderdi batzuk. Nabaria da fisika orokorreko terminoen doitasuna, eta bioteknologiako terminoen doitasun apala. Fisikaren kasuan, atomo- eta partikula-fisikaren arloko terminoekin batera fisika orokorreko termino ugari agertzearen esplikazioa izan liteke testu gehienak ez izatea espezialisten arteko komunikazioak, ikasmaterialak eta dibulgazioa baizik.

Corpusaren tamainaren eragina ere ageri da emaitzetan. Corpus handienarekin, informatikakoarekin, lortu dira doitasun-emaitza onenak. Baina bioteknologiako corpusa atomo- eta partikula-fisikakoaren bikoitza da, eta kasu horretan ez da nabari tamaina handiagoaren eragina. Emaitzen arteko aldea, gure ustez, bioteknologiaren hedadura handiagoari eta teknologia aplikatuek bertan duten pisu handiagoari egotzi behar zaie; izan ere, faktore horiek dira, erregistroarekin batera seguru aski, determinatuko dutenak zenbateraino agertuko diren corpusean jakintza-arlo desberdinetako terminoak.

4.4.3. Estaldura

Web-corpusen estaldura neurtzea interesgarria da hiztegian dauden terminoek erabilera errealean duten presentzia egiaztatzeko.



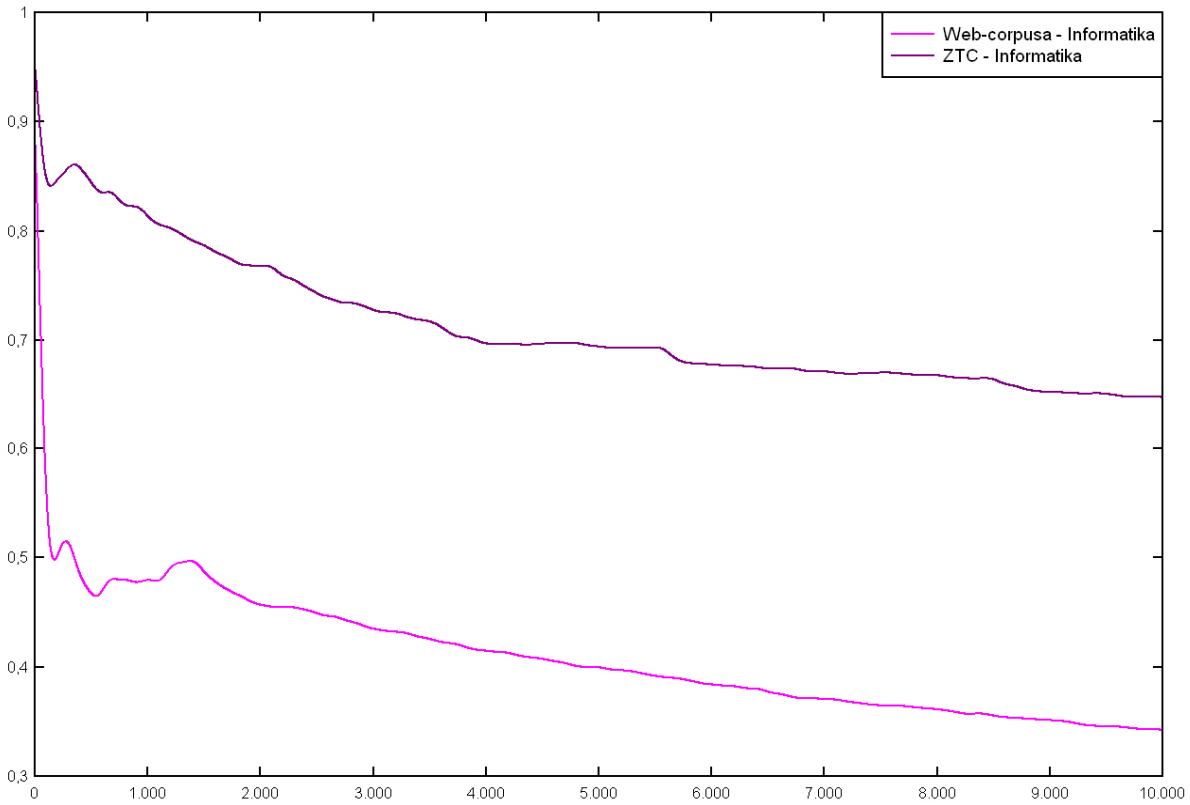
4. irudia. Termino-erauzketen arlo-estaldura ZTHrekiko (baliozko termino guztiekiko).

Begi-bistakoa da emaitza onenak atomo- eta partikula-fisikaren corpusarenak direla, eta okerrenak bioteknologiaren corpusarenak. Irudi luke hiru web-corpusak, batez ere informatikakoa eta bioteknologiakoa, ez direla adierazgarriak. Baina horrek ez du esan nahi ezinbestean corpusa eratzeko tresnaren hutsegitea denik. Esaterako, atomo- eta partikula-fisikaren kasuan, ZTH hiztegian dauden 474 terminoetatik, 150 ez daude termino erauzien artean; horiek Interneten benetan badiren egiaztatu dugu, eta 42 ez dira Googleren emaitzetan ageri; Interneten diren gainerakoetatik, 4 ez dira ageri corpuseko testuetan, eta beste 104ak ez

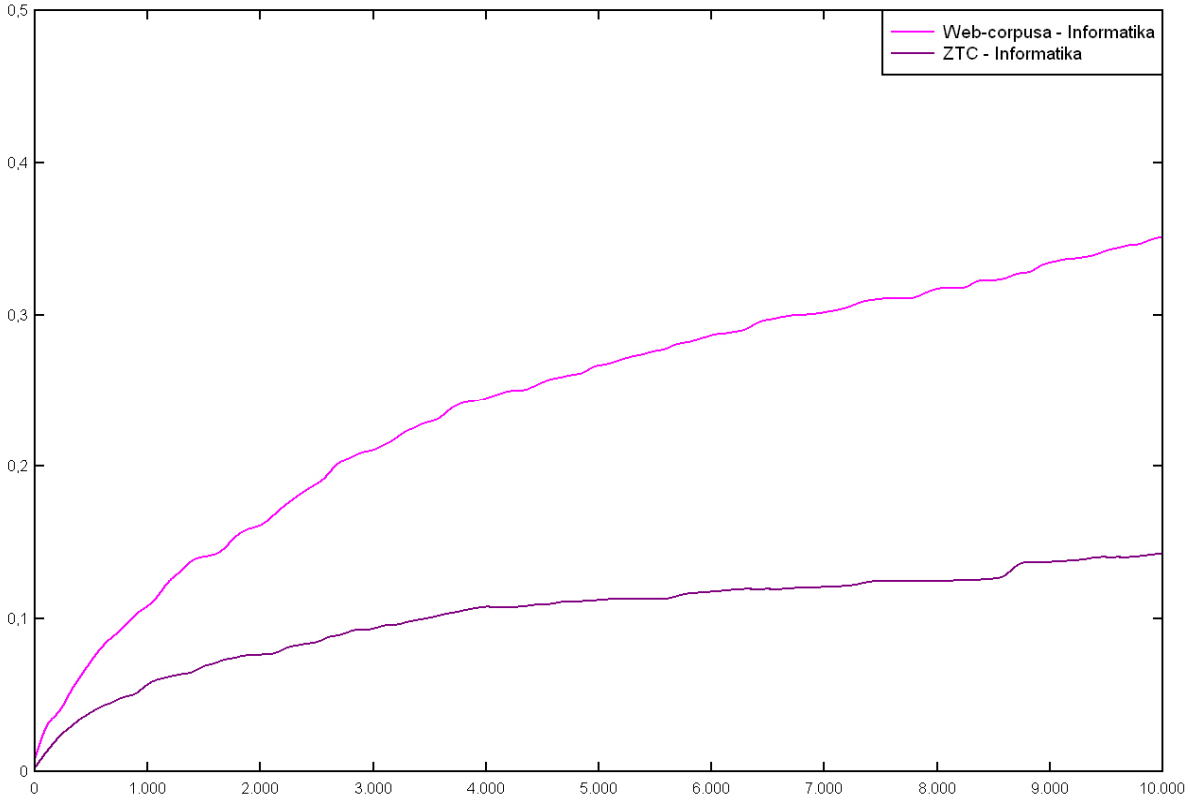
ditu *Erauzterm*ek erauzi, baina horietatik 101 termino behin baino ez dira corpusean ageri, eta hori oso maiztasun txikia da termino-erauzle batentzat. Beraz, euskarazko webaren estaldura arazo bat da. Horiek horrela, ZTHn parte hartu duten adituek hautatu dituzten kontzeptu batzuen euskarazko terminoak ez dira Interneten ageri, edo maiztasun txikia dute, eta horrek, oraingoz behintzat, nolabaiteko muga jar diezaioke Internet espezializazio-arlo batzuetan datu-iturri bakartzat erabiltzeko alternatibari.

4.4.4. ZTCrekiko konparazioa

Nolakoak dira emaitza horiek webekoa ez den corpus batekin konparatuta? *Zientzia eta Teknologiaren Corpusa*ren informatika-arloko azpicorpusa (332.745 hitz) era berean prozesatu dugu, eta 5. eta 6. irudietan daude haren doitasun- eta estaldura-emaitzak, informatikako web-corpusarekin konparatuta.



5. irudia. Termino-erauzketen arlo-doitasuna (baliozko termino guztiekiko): informatikako web-corpusa eta ZTCko informatika-azpicorpusa.



6. irudia. Termino-erauzketen arlo-estaldura ZTHrekiko (baliozko termino guztiekiko): informatikako web-corpusa eta ZTCko informatika-azpicorpusa.

Deigarria da doitasun hobea ZTCrekin lortu izana, baina estaldura hobea web-corpusarekin. Nola interpretatu emaitza hori? Gure iritziz, ZTCko informatika-arloko azpicorpusa "espezializatuagoa" da (eskuz sailkatutako testuez osatua da), eta horregatik erauzketaren emaitzetan informatikako terminoen ehunekoa handiagoa. Baina web-corpusa handiagoa da, eguneratuagoa, eta, hain zehatza ez bada ere, ZTHrako aukeratu diren informatikako termino gehiago daude bertan.

5. Ondorioak eta gerorako ideiak

Internetek corpusgintzarako eskaintzen bide dituen aukerak erakargarriak dira edozein hizkuntzarentzat, eta are garbiago baliabide urrikoentzat. Nolanahi ere, aukera horiek benetakoak eta bideragarriak direlako hipotesia egiaztatu egin behar da.

Aurkeztu ditugun esperimentuetan, baieztatu dugu Internetetik Co3 tresnaren bidez automatikoki eratu ditugun corpusak espezializatuak direla, hein desberdinean bada ere. Doitasun- eta estaldura-emaitzek iradokitzen dute arloaren hedadurak, corpusaren tamainak eta zientzia/teknologia izaerak eragina duela bertatik erauzten diren terminoetan eta horien arlokako banaketan. Doitasuna handiena corpus handienarekin lortu da; estaldura onena, berriz, zientzia-arlo erlatiboki ez oso zabal batekin. Dena den, ikusi dugu estalduraren emaitzetan eragin handiagoa duela euskarazko webaren beraren estaldurak (arlo eta genero batzuetako dokumentuak gutxi izateak), corpus-tresnaren eta termino-erauzlearen estaldurek baino. Bestetik, ZTCrekin egindako konparazioak pentsarazten digu webetik eratutako corpusa eguneratuagoa dela, eta, doitasun txikiagoz bada ere, termino gehiago erauz daitezkeela.

Horrenbestez, gure ondorioa da merezi duela Internet terminologia-lanerako corpus-iturritzat erabiltzen hastea, eta garatu ditugun tresnek bideragarria egiten dutela estrategia hori. Lortutako hitz-kopuruaren eta kostuaren arteko erlazioa dela eta, abantailak nabariak dira. Dena den, arlo eta genero batzuetan euskarazko webak oraindik duen estaldura-arazoa kontuan izanik, ezin da, oraingoz behintzat, iturri bakartzat erabili, eta inprimatutako obren corpusekin konbinatu behar litzateke.

Web-corpusak eratzeko teknika eta tresnen aldetik, uste dugu ikertze-ildo nagusiak liratekeela hazi-dokumentuen hautaketa hobetzea, hazi-hitzen artean hitz anitzeko terminoak ere kontuan hartzea, eta espezializazio-arloaren izaera eta hedaduraren eragina aztertzea.

Laburbilduz, uste dugu Internet corpusgintzan erabiltzeko lehen urratsak egiten hasiak garela, eta bide horretan aurrera egin behar genukeela datozen urteetan.

Bibliografia

ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. & URIZAR, R. (2004): "An Xml-Based Term Extraction Tool for Basque", *in Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*. Lisboa: ELRA: 1733-1736 or.

ARETA N., GURRUTXAGA A., LETURIA I., ALEGRIA I., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N. & SOLOGAISTOA A. (2007): "ZT Corpus: Annotation and tools for Basque corpora", *in Proceedings of Corpus Linguistics,* Birmingham: Birminghamgo Unibertsitatea.

BARONI, M.; BERNARDINI, S. (2004): "BootCaT: Bootstrapping corpora and terms from the web", *in Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*, Lisboa: ELRA: 1313-1316 or.

BIBER, D.. (1993): "Representativeness in Corpus Design", *Literary & Linguistic Computing* 8. 243-257. or.

CABRÉ, M.T. (2001): "Consecuencias metodológicas la nueva propuesta teórica (I)", *in* CABRÉ, M.T *et al*. (Ed.): *La Terminología científico-técnica: reconocimiento, análisis y extracción formal y semántica*, 27-36 or., Bartzelona: IULA-Institut Universitari de Lingüística Aplicada-Universitat Pompeu Fabra.

FLETCHER, W.H. (2004): "Making the web more useful as a source for linguistic corpora", *in* U. Connor and T. Upton (Ed.) *Corpus Linguistics in North America 2002*, Amsterdam: Rodopi: 191-205 or.

KEHOE, A. & RENOUF A. (2002): "WebCorp: Applying the Web to Linguistics and Linguistics to the Web", *Proceedings of the WWW2002 Conference*. Honolulu: W3C.

KILGARRIFF, A. & GREFENSTETTE, G. (2004): "Introduction to the special issue on the Web as corpus", *Computational Linguistics*, 29: 333-348 or.

KILGARRIFF, A. (2007): "Googleology is Bad Science", *Computational Linguistics* 33, 1, 147-151 or.

LETURIA, I., GURRUTXAGA, A., ALEGRIA I. and EZEIZA A. (2007): "CorpEus, a web as corpus tool designed for the agglutinative nature of Basque", *Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve: Presses Universitaires de Louvain: 69-81 or.

LETURIA, I., SAN VICENTE, I., SARALEGI. X. (2009): "Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet", *Proceedings of 5th International Web as Corpus Workshop (WAC5)*. Donostia, Spain: ACL-SIGWAC, 53-61 or.

LÜDELING, A., EVERT, S. & BARONI, M. (2007): "Using Web data for linguistic purposes", *in* HUNDT, M., NESSELHAUF, N & and BIEWER, C. (Ed.) *Corpus Linguistics and the Web,* Amsterdam: *Language & Computers* 59. Rodopi: 7-24 or.

POMIKÁLEK, J., RYCHLÝ, P. & KILGARRIFF, A. (2009): "Scaling to Billion-plus Word Corpora", *Advances in Computational Linguistics*, Mexiko: Instituto Politécnico Nacional 41 liburukia, 2009, 3-13 or.

RENOUF, A. (2007): "Corpus development 25 years on: from super-corpus to cyber-corpus", *in* FACCHINETI, R. (Ed.) *Corpus linguistics 25+ years on*, Rodopi: Amsterdam - New York.

SARALEGI, X. & LETURIA, I. (2007): "Kimatu, a tool for cleaning non-content text parts from HTML docs", *Proceedings of the 3rd Web as Corpus workshop*, Louvain-la-Neuve, Belgium: Presses universitaires de Louvain, 163-167 or.

SARALEGI, X. SAN VICENTE, I. & GURRUTXAGA, A. (2008): "Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain", *Proceedings of Building and using Comparable Corpora workshop*, Marrakech, Maroko: ELRA.

SINCLAIR, J. (2005): "Corpus and Text - Basic Principles", *in* WYNNE, M. (Ed.) *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books: 1-16 or. On line: http://ahds.ac.uk/linguistic-corpora/ [2010-03-29].

AURKIBIDEA