CORPUSEN IRAULTZA: MAIZTASUNEN KORAPILOA

Josu Landa
(Ametzagaiña, Lasarte-Oria)

AURKIBIDEA

0. Sarrera

Azken urteotan, Internet bidezko informazioaren erabilerak ekarri duen iraultzetako bat corpusena izan da. Gero eta ugariagoak dira corpusak, gero eta anitzagoak. Weba bera corpus erraldoia da, baina etengabe sortzen ari dira bestelako corpus berezituak ere. Erabiltzaile arrunten eremu zabalean ez ezik, ikerketaren alorrean gero eta garrantzi handiagoa du aurrez egindako testuekin kontsultak eta bestelako aplikazioak egin ahal izateak. Duela urte batzuk batik bat hizkuntzaren inguruko ikerleen kasua baldin bazen, gaur egun gero eta zientzi adar gehiagotan gertatzen da corpusen erabilera hori. Arkimedes parafraseatuz, "emadazue corpus bat eta mundua mugituko dut" izan zatekeen artikulu honen idazpurua.

Guri interesatzen zaigunera makurtuta, komeni da ikerketaren alorrean corpusez egiten den erabilera zein den aztertzea. Corpusak aztertzeko, eta haietatik baliagarriak gerta dakizkigun datuak ondorioztatzeko, erreminta ezinbesteko eta indar handikoa daukagu: estatistika. Haren bitartez lor ditzakegun datuak, gero aztertu eta ondorioak ateratzeko balia ditzakegu, baita etorkizuneko sorkuntza lanetan berrerabiltzeko ere.

1. Lexikoaz harago

2000az geroztik, Ametzagaiña lankidetzan ari da EHUko Euskara Institutuarekin, eta hortik sortu dira corpusetan oinarritutako hainbat aplikazio, guztiak Institutuaren beraren web orrian erabiltzeko moduan direnak[1]:

— erreferentziako corpusa:
Ereduzko Prosa Gaur (EPG)
— EPGtik egindako erauzketa lexikoa:
Hiztegi Batua Euskal Prosan (HBEP)
— OEHren eta EPGren arteko corpus alderatzea:
Lexikoa Atzo eta Gaur (LAG)
— corpusean oinarritutako hiztegia:
Egungo Euskararen Hiztegia (EEH)
— corpus dinamikoa:
Ereduzko Prosa Dinamikoa (EPD)

Orain arte, corpusei egindako ustiapena batez ere lexikoari edo terminologiari begirakoa izan da, eta alde horretatik goian aipatutako aplikazio horiek ez dira salbuespena.

[1] http://www.ei.ehu.es/

Askoz murritzagoak dira lexikografiaz haragoko esparruetan egin izan diren saioak:

- sintaxia
- fraseologia
- diskurtsoaren eraikuntza

Apurka bada ere, corpusetatik erauzitako datuak beste esparru horietara ere zabaltzen hasi behar litzakeela pentsatzen dugu, eta horretara gatoz gaur, oso modu apalean bada ere.

2. Hitz multzoen maiztasunak

Hitz multzoen eta haien aldagarritasuna neurtzeko saioa da hemen aurkeztu nahi duguna. Xedea hiru hitzeko multzoen maiztasunak neurtzea eta alderatzea da. Corpusean ageri diren hitz multzo horien maiztasunak aztertuz gero, eta azpicorpusen arteko desberdintasunak behatuta, perpausak osatzeko orduan ageri den aldagarritasunari buruzko ahalik eta informazio handiena lortu ahal izango dugu.

Abiapuntua *Ereduzko Prosa Gaur* corpusa izan da, 25 bat milioi testu-hitzez osatua. Horko hitz-hirukote guztiak erauzi dira, nahiz eta bazter utzi diren honako osagairen bat zituztenak:

— puntuazio markak
— entitate izenak
— zifrak

Denera, 15 bat milioi hitz-hirukote lortu dira horrela.

2.1. Hirukote erabilienak

Hauek dira gehien erabiltzen diren hiru hitzeko multzoak:

2954	hain zuzen ere
1597	hori dela eta
1497	behin eta berriz
1485	ez dut uste
1432	hori ez da
1403	egin behar izan
1381	behar izan zuen
1312	argi eta garbi
1311	behin eta berriro
1293	eta ez da
1240	behin baino gehiagotan
1206	baina ez da
1107	esan nahi du
1089	nahi izan zuen
1084	bertan behera utzi
1041	dela esan zuen
1024	hala eta guztiz
1021	bat besterik ez
1016	bat egin zuen
988	izan behar du

955 besterik ez da
954 horrek ez du
951 egin behar da
940 bat baino gehiago
919 baina ez zuen

Ezkerreko zutabean agerpen kopurua ageri da.

Zerrenda horretan, nolabaiteko bereizketa egin genezake: alde batetik, tartean aditz laguntzaileren bat edo partizipioren bat dutenak (ezkerreko zutabean), eta gainerakoak (eskuineko zutabean).

1597 hori dela eta
1485 ez dut uste
1432 hori ez da
1403 egin behar izan
1381 behar izan zuen
1293 eta ez da
1206 baina ez da
1107 esan nahi du
1089 nahi izan zuen
1041 dela esan zuen
1016 bat egin zuen
988 izan behar du
955 besterik ez da
954 horrek ez du
951 egin behar da
919 baina ez zuen

2954 hain zuzen ere
1497 behin eta berriz
1312 argi eta garbi
1311 behin eta berriro
1240 behin baino gehiagotan
1084 bertan behera utzi
1024 hala eta guztiz
1021 bat besterik ez
940 bat baino gehiago

Pentsatzen dugu eskuineko zutabeko multzoa dela arreta gehien merezi duena, horrek adieraz dezakeelako diskurtsoa eraikitzeko orduan euskararen egungo erabilerak dituen makur edo herren batzuk zeintzuk diren:

- Alde batetik, adabaki erretoriko dei genitzakeenak ageri dira. Diferentzia handiz lehen postuan den *hain zuzen ere* nonahikoa dugu kasurik paradigmatikoena.
- Deigarria da, halaber, aldizkotasuna edo kopurua adierazten duten multzoen ugaritasuna: *behin eta berriz, behin eta berriro, behin baino gehiagotan, bat besterik ez, bat baino gehiago*.
- Azkenik, hitz bakarreko lemen sinonimoak daude. *Bertan behera utzi* gehiegi erabiltzen ote den neurtzeko, aintzat har bedi berorren agerpenetatik %80a prentsan ageri direla, eta ia beti *ekitaldia, legea* eta antzeko hitzekin lotuta.

Maiztasun handieneko 25 multzoak ez ezik, zerrenda luzeagoak aztertuz gero, ondorio osoagoak eta zehatzagoak atera ahal izango lirateke. Dudarik ez dago, ondorioztatze horiek

egiteko, informatikariok ez beste alor batzuetako adituek dutela hitza. Haiek jakingo zer irizpide erabili datuok aztertzeko orduan.

2.2. Azpicorpusetako emaitzak

Ikus dezagun, orain, corpusa osatzen duten testuen iturrien arabera emaitzak nola aldatzen diren.

2.2.1. Prentsa vs. liburuak

(Oharra: lehen zutabean, azpicorpuseko maiztasun postua dago; bigarrenean, corpus osoko maiztasun postua; eta hirugarrenean, bien arteko aldea).

PRENTSA				LIBURUAK			
1	1	(0)	hain zuzen ere	1	1	(0)	hain zuzen ere
2	2	(0)	hori dela eta	2	3	(1)	behin eta berriz
3	16	(13)	dela esan zuen	3	9	(6)	behin eta berriro
4	27	(23)	esan zuen atzo	4	18	(14)	bat besterik ez
5	15	(10)	bertan behera utzi	5	7	(2)	behar izan zuen
6	5	(-1)	hori ez da	6	11	(5)	behin baino gehiagotan
7	41	(34)	joan den astean	7	6	(-1)	egin behar izan
8	47	(39)	adierazi zuen atzo	8	19	(11)	bat egin zuen
9	6	(-3)	egin behar izan	9	4	(-5)	ez dut uste
10	23	(13)	egin behar da	10	14	(4)	nahi izan zuen
11	8	(-3)	argi eta garbi	11	13	(2)	esan nahi du
12	10	(-2)	eta ez da	12	8	(-4)	argi eta garbi
13	4	(-9)	ez dut uste	13	21	(8)	besterik ez da
14	57	(43)	egin zuen atzo	14	10	(-4)	eta ez da
15	12	(-3)	baina ez da	15	31	(16)	alde egin zuen

Desberdintasun handia dago hirukoteen zerrendetan batetik bestera, eta hori berez aski adierazgarria dela iruditzen zaigu. Kontuan hartuta azpicorpus horietako bakoitza 12,5 bat milioi testu-hitzez osatua dela, espero izatekoa zen kopuruak berak berdintzera eramango zituela biak. Baina errealitateak beste zerbait erakusten digu, prentsako testuen emaitza aski berezitua baita liburuenetik. Hedabideen ofizioari lotutako adierazpideak asko markatzen du prentsako testuetatik erauzitako zerrenda, denbora erreferentzien aldetik (*atzo* hitzaren ugaritasuna kasu) edo komunikazio aditzen aldetik (*esan, adierazi*).

2.2.2. Jatorrizko liburuak vs. itzulpenak

JATORRIZKOAK				ITZULPENAK			
1	1	(0)	hain zuzen ere	1	1	(0)	hain zuzen ere
2	3	(1)	behin eta berriz	2	55	(53)	oihu egin zuen
3	11	(8)	behin baino gehiagotan	3	18	(15)	bat besterik ez
4	9	(5)	behin eta berriro	4	13	(9)	esan nahi du
5	6	(1)	egin behar izan	5	21	(16)	besterik ez da
6	4	(-2)	ez dut uste	6	7	(1)	behar izan zuen
7	14	(7)	nahi izan zuen	7	9	(2)	behin eta berriro
8	7	(-1)	behar izan zuen	8	3	(-5)	behin eta berriz

9	19	(10)	bat egin zuen
10	18	(8)	bat besterik ez
11	39	(28)	bat edo beste
12	43	(31)	zer edo zer
13	44	(31)	ez dakit zer
14	38	(24)	baina ez zen
15	33	(18)	eta ez zuen

9	8	(-1)	argi eta garbi
10	19	(9)	bat egin zuen
11	65	(54)	bat baizik ez
12	17	(5)	hala eta guztiz
13	25	(12)	baina ez zuen
14	31	(17)	alde egin zuen
15	4	(-11)	ez dut uste

Beste mota bateko desberdintasuna azaltzen zaigu, tratatutako azpicorpusak domeinu bertsukoak direnean. Pentsa liteke hizkuntza-erregistro aldetik ez dagoela aparteko bestelakotasunik euskaraz idatzitako liburu baten eta erdaratik itzulitako liburu baten artean. Baina, bi zerrenden erkatzeak kontrako ustera hurbiltzen gaitu. Badirudi sorkuntzak --itzulpenaren aldean-- perpausa eraikitzeko modu desberdin bat eragiten duela.

2.3. Multzo erabilienen pisua

Lehenago esan bezala, ez da gure lana erauzitako zerrendak aztertzea eta hizkuntzalariei legozkiekeen ondorioak ateratzea. Baina bada beste hurbiltze matematiko bat gure eskueran dagoena: azpicorpus bakoitzaren hitz-multzo erabilienekin estaldura indizeak kalkulatzea, bi eratako konparazioak eginda: corpus osoarekin eta azpicorpusarekin berarekin.

2.3.1. Corpus osoarekiko

Erreferentziako corpus osoa nolabaiteko estandar baten zehaztapen bezala kontsideratuz gero, corpus horren barruko azpicorpusetako hitz multzoen erabilera aztertu eta kopuru erlatiboak neurtuta, azpicorpus horiek estandar horrekin duten distantzia neurtzeko balio diezaguke. Alegia, noraino den estandarrarekiko urrun edo hurbil.

Horretarako, corpus osoko hirukote erabilienak hartu dira, eta corpus osoaren zenbatekoa suposatzen duten neurtu. Hiru eskala desberdin erabili dira: 10 hirukote erabilienak, 15 eta 20.

Adibidez, corpus osorako, 10 hirukote erabilienek corpus osoaren 9,89ko estaldura ageri dute (10.000koetan adierazita), 15 erabilienek 13,51 eta 20 erabilienek 19,70.

Corpus osoko hirukote horiek berek, azpicorpus bakoitzean, zenbateko estaldura suposatzen duten kalkulatuz gero, ondoko taulako emaitzak lortzen ditugu:

	10	15	25
EPG	9,89	13,51	19,70
prentsa	9,84	13,40	19,49
liburuak	9,59	13,21	19,44
Berria	10,94	15,13	22,16
Herria	3,17	3,58	4,84
Berria – Ekonomia	8,77	12,68	16,81
Berria – Euskal Herria	12,04	16,39	25,95

Berria – Gaiak	8,43	11,45	18,10
Berria – Harian	10,88	15,86	24,22
Berria – Kirolak	11,82	15,83	22,67
Berria – Kultura	10,01	13,85	18,80
Berria – Mundua	10,27	14,91	19,87
liburuak – Jatorrizkoak	9,98	13,72	19,46
liburuak – Itzulpenak	9,53	13,20	20,66
liburuak – Mendebaldea	10,22	14,06	20,69
liburuak – Ekialdea	3,32	5,08	8,26

Lehen ondorioa da, 10, 15 edo 20 hirukote hartuta, ez dagoela desberdintasun haien artean nabarmenik; hau da, metodologikoki, aski izan daiteke erreferentzia eskala horietako bakarra hartzea (adibidez, 25ekoa). Pentsa liteke, era berean, 25etik gora hartuz gero, azpicorpusen arteko desberdintasunak mantendu egingo liratekeela, gutxi gorabehera.

Datuak xehe aztertuz gero, beste hainbat ondorio erator ditzakegu (aintzat hartuko diren zifra guztiak azken zutabeari dagozkio, hau da, 25 multzo erabilienen datuei):

- Liburuetako eta prentsako testuen artean ez dago bereizgarri nabarmenik.
- Aurrez pentsatzekoa zen bezala, *Herria* astekariko testuak dira estandarretik gehien urruntzen direnak, 25 multzo erabilienek corpus osoaren 4,84 baino ez baitute hartzen *(Berria*k, aldiz, 22,16).
- *Berria*ren barruan, "Euskal Herria" saila da, nolabait esanda, errepikakorrena (25,95). Aldiz, barietate handiena "Ekonomia" sailean aurkitzen dugu (15,81).
- Liburuen azpicorpusean ez dago desberdintasunik, itzulpenak izan edo jatorriz euskaraz idatziak izan. Beraz, itzulpenaren ekintza (abiapuntu bezala beste hizkuntza bat hartzea) ez da aldagai adierazgarria aztertzen ari garen gai honetarako.

Liburuen azpicorpusaren barruan, idazle edo itzultzaile bakoitza ere azpicorpus bezala jo daiteke. Mekanismo bera azpicorpus murritzago horiei aplikatu diegu. Horretarako, corpusean 4 liburu edo gehiago dituzten idazle eta itzultzaileak erauzi dira. Hona neurketa horren emaitzak:

	10	15	25
Aristi P	6,06	8,94	17,26
Borda I	1,03	1,76	3,32
Cano H	7,77	11,59	19,72

Igerabide JK	10,67	14,55	25,23
Irigoien JM	9,46	13,31	17,21
Izagirre K	5,52	7,71	14,28
Jimenez E	11,40	15,27	19,76
Lertxundi A	6,31	8,95	13,73
Mendiguren X	11,32	14,53	20,77
Zaldua I	18,40	25,67	36,19
Garzia J	7,80	11,22	15,08
Mendiguren I	9,45	12,24	17,94
Muñoz J	10,67	14,15	21,09
Navarro K	8,59	11,80	18,31
Olarra X	10,86	13,79	19,82
Rey F	6,68	8,55	15,41
Zabaleta J	11,50	16,25	24,18

- Azpicorpusa txikiagoa den neurrian, estandarrarekiko distantzia handitu egiten da.
- Aldakortasun handia ageri da idazleen artean, eta diferentzia handia dago multzo gehien (36,19) eta gutxien (3,32) errepikatzen diren azpicorpusen artean.
- Itzulpenetan aldakortasuna askoz txikiagoa da, jatorrizko liburuetan baino. Nabarmen, gainera.

2.3.2. Multzo erabilienen pisua azpicorpus berean

Aurreko atalean, azpicorpusek estandarrarekin duten distantzia neurtu badugu, oraingo honetan beste neurketa modu bat da proposatzen dena. Azpicorpus bereko hitz multzo erabilienak hartzen dira soilik kontuan. Erreferentzia, beraz, azpicorpusa bera da. Azpicorpus horretan gehien erabiltzen diren hirukoteek, osoaren zenbatekoa suposatzen duten neurtu nahi da. Azpicorpus bakoitzaren aldakortasuna eta errepikapena neurtzeko erabil daitezke eratorritako datuok. 10 hirukote erabilienak hartuz gero, hauexek dira zifrak:

EPG	9,89
prentsa	11,80
liburuak	10,52
Berria	13,54

Herria	17,26
Berria – Ekonomia	23,19
Berria – Euskal Herria	19,66
Berria – Gaiak	30,87
Berria – Harian	13,80
Berria – Kirolak	15,25
Berria – Kultura	14,93
Berria – Mundua	31,06
liburuak – Jatorrizkoak	11,15
liburuak – Itzulpenak	12,30
liburuak – Mendebaldea	11,31
liburuak – Ekialdea	14,12

Eta hauek, idazle-itzultzaileak banaka bereiziz gero:

Aristi P	16,44
Borda I	11,43
Cano H	16,97
Igerabide JK	29,96
Irigoien JM	24,30
Izagirre K	18,85
Jimenez E	25,18
Lertxundi A	17,47
Mendiguren X	21,84
Zaldua I	40,22
Garzia J	12,97
Mendiguren I	65,49
Muñoz J	25,84
Navarro K	29,42
Olarra X	19,42
Rey F	38,13
Zabaleta J	28,28

Estandarrarekin distantzia neurtzerakoan (2.3.1 atala) ez bezala, oraingoan batetik besterako aldakortasuna erabatekoa da. Diferentziak oso handiak dira, eta errepikakortasun indizerik altuena 65,49ra iristen da.

Esperimentua zuzena izango balitz, esan nahiko luke indize bakoitzak balukeela zerikusirik azpicorpus bakoitzaren testuek diskurtsoa eraikitzeko duten barietatean eta egitura aniztasunean. Eta zergatik ez, irakurgarritasuna edo ulergarritasuna bezalako faktoreetan.

3. Ondorioak eta geroko lanak

Aurren-aurrena, esan behar dugu aurkeztu dugun honek lehen hurbiltzea besterik ez zuela izan nahi, corpusetan hitzaz eta lexikoaz haratago egin daitezkeen azterketen aitzin-pausoa. Esperimentu honetan maiztasunak tratatzeko erabilitako sistema izugarri primitiboa izan da:

- Hiru hitzeko multzoak soilik hartu dira aintzat.
- Puntuazio markak bazter utzi dira, eta horrek asko erlatibizatzen du emaitza, bistan denez puntuazioa oinarri-oinarrizko elementua baita fraseologian.
- Ez da inolako informazio linguistikorik erabili, ez lexikorik (hitzak, eta ez lemak) ez morfosintaktikorik (kategoria gramatikala, e.a.).

Nolanahi ere, hain lanabes traketsak erabilita ere, uste dugu saioaren emaitza positiboa izan dela, hots, perpausak osatzeko orduan erabiltzen diren hitz-kateatzeei buruz, eta haien aldagarritasunari buruz, aski datu interesanteak ondorioztatu ahal izan direla.

Diskurtsoan edo estiloan zer detektatu nahi den zehaztuz gero, berariaz egokitutako tresnak sor daitezke, beste hainbat zertzelada eta mekanismo aintzat hartuko dituztenak:

- Lematizatzaileak ematen duen informazioa.
- Hirutik beherako edo gorako hitz multzoak.
- Hitz hurrenkera hutsaz gain, patroi morfosintaktikoak erabiltzea.

AURKIBIDEA