BEGIRATU BAT CORPUS-BALIABIDEEI

Nerea Areta
nereaa@elhuyar.com
Antton Gurrutxaga
agurrutxaga@elhuyar.com
Igor Leturia
igor@elhuyar.com
Elhuyar Fundazioa
Hizkuntza Zerbitzuak - I+G+b

1 Sarrera

Maiz esan ohi da hizkuntza batek hizkuntza-teknologien garapenean duen maila lotuta dagoela hizkuntza horren 'osasunarekin', horren adierazle izateraino. Ez da aztia izan behar bi alderdi horien arteko lotura badagoela sumatzeko. Mundu zabalean posizio 'indartsuan' dauden hizkuntzak dira hizkuntza-teknologietan, eta, horien artean, corpusgintzan, baliabide gehien dituztenak. Horrek, berriz, zerikusi zuzena du hizkuntza horiek mintzatzen diren herrialdeen botere ekonomikoarekin ere, zientzia- eta teknologia-ikerkuntzan inbertitzeko baliabide handiak, eta gizarte teknologizatuak, behar baitira, produktuak eta zerbitzuak merkaturatzea bideragarria izango bada.

Sintoma hor dago. Besterik da, ordea, baldintza beharrezkoa eta nahikoa den, hau da, corpusgintzan oso garatuta ez dauden hizkuntzak etorkizun beltza izatera kondenatuta dauden, edo garapen handiak etorkizuna bermatzen duen.

Horretarako, delako 'osasun' hori zer den, zein adierazleren bidez neur daitekeen zehaztu behar da. Osasunarekin erlazionatu ohi diren adierazle batzuk aipatzearren: transmisioa, gizarteko erregistro eta arloetan zenbateraino erabiltzen den, hezkuntzan, komunikabideetan edo lan-jardueran duen presentzia, hiztun-kopurua, argitalpen-kopuruak, literatura, Interneteko eduki-kantitatea…

Gu ez gara soziolinguistak, eta nekez egin genezake ekarpenik hain espezializatua den ikertze-arlo horretan. Artikulu honetan, gure asmo bakarra da euskarazko eta beste hizkuntza batzuetako corpus-baliabideetan eta corpusgintzan egin dugun datu-bilketa aurkeztea, eta datu horien azterketan oinarrituta gogoeta xume bat egitea, non gauden eta norantz jo behar genukeen hausnartzeko.

2 Corpusa eta corpus-motak

Corpusei buruzko artikuluen hasieran, ohikoa da *corpus* kontzeptua definitzea, eta corpus-motak bereiztea. Corpus-tipologia onartu eta estandar samarra dago gaur egun, eta euskaraz ere azaldua da literaturan (IXA taldea & Elhuyar Fundazioa, 2007). Hemen ez dugu horretan sakonduko, baina komeni da oinarrizko ideia batzuk ematea.

Corpus kontzeptuaren oinarrizko ideia da hizkuntza-erabilera errealekoak diren datuen bilduma dela. Ikuspegi zabal, beharbada zabalegi, horren arabera, edozein testu-bilduma izan liteke corpus edo erabil liteke corpustzat, baina, normalean erabiltzen den kontzeptua jasoko badugu, honakoak gehitu behar genituzke: hizkuntza aztertzeko edo hizkuntzari buruzko datuak eskuratzeko asmoz bildua edo horretarako erabilia izatea, nolabaiteko diseinu batez edo irizpide batzuez eratua, (gaur egun) formatu elektronikoan bildua, eta linguistikoki prozesatua. Ezaugarri edo aldagai horien balio zehatzen arabera antolatzen da corpus-tipologia: erreferentzia-corpusa/corpus berezi edo espezializatua, idatzizko testuen corpusa/hizketa- edo ahozko corpusa/corpus multimodala, corpus orekatua/oportunista, corpus elebakarra/elebiduna/eleaniztuna eta, azken bi horien barruan, corpus paraleloak/konparagarriak…

3 Datu pare bat corpusgintzaren historiaz

1964 jo ohi da, *Brown* corpusa argitaratu zen urtea hain zuzen ere, corpusgintza modernoaren eta corpus-hizkuntzalaritzaren hasiera-urtetzat. *Brown* corpusa izan zen lehen corpus 'elektronikoa'. Garai haietan, teoria sortzaile-bihurtzailea zen nagusi hizkuntzalaritzan. N. Chomsky-k 1957an argitaratu zuen *Syntactic Structures* obrak linguistikaren paisaia erabat aldatua zuen ordurako. Sortzaileentzat, corpusa ez zen aintzat hartzeko moduko baliabidea[1], eta hori erabiltzearen kontrako jarrera nabaria zen. Ulertzekoak dira orduan J. Svartvik-en hitz hauek giro hura gogoratzen duenean:

“There might have been moments when being so named [a corpus linguist] felt like discovering your name on the passenger list for the Titanic” (Svartvik 2007).

Gauzak asko aldatu dira harez geroztik. Eztabaida ez da agortua, baina garai bateko jarrera itxiak samurtuz joan dira, eta gaur egun inork ez du zalantzan jartzen, zein ere teoria linguistikoren aldekoa den, corpusak beharrezkoak direla hizkuntza ‘den bezala’ aztertu nahi bada (Rojo 2002). Corpusak, gainera, ez dira hizkuntzalaritzaren arloko ikerkuntzan soilik erabili, hiztegigintzan (batez ere), hizkuntza-irakaskuntzan, hizkuntza-teknologietan (teknika estatistikoak aplikatzeko, itzulpen automatikoan, hizketaren tratamenduan…) eta beste zenbait arlotan ere bai.

Horiek denak egiteko ordea, corpusak eurak behar dira. Corpusgintzan izan den bilakaera aztertzea interesgarria izan daiteke, gure ustez, euskara non dagoen eta etorkizunean nora begira jarri behar genukeen pentsatzen hasteko. Ingelesezko corpusen historian, honako aldiak bereizi ditu A. Renouf-ek (Renouf 2007):

1960s onwards	The one million word (or less) Small Corpus • standard • general and specialised • sampled • multi-modal, multi-dimensional
1980s onwards	The multi-million word Large Corpus • standard • general and specialised • sampled • multi-modal, multi-dimensional
1990s onwards	The 'Modern Diachronic' Corpus • dynamic, open-ended, chronological data flow
1998 onwards	The Web as corpus • dynamic, open-ended, chronological data flow
2005 onwards	• the Grid; pathway to distributed corpora • consolidation of existing corpus types

Aurreko taulak ez du bere horretan beste hizkuntzetarako balio, batez ere aldien arteko mugak ez direlako garai horietan gertatu, edo hizkuntza batzuetan urrats batzuk hasi ere ez direlako egin. Dena den, corpusgintzaren historian eragin handia izan du ingelesak (aitzindari, gehienetan), eta bilakabidearen eta joeren eskema orokortzat har genezakeela uste dugu.

4 Corpus-baliabideen analisia

Corpusen mundura hurbiltzen dena berehala konturatuko da zein mundu zabala den. Era askotako corpusak daude, helburu askotarako eratuak eta erabiliak. Gehienetan, hizkuntzalaritzaren arlokoak izaten dira helburuok, baina, sarritan, hizkuntzalaritzaz kanpoko eremuak ere ukitzen eta jorratzen dira, soziolinguistikoak nagusiki.

Halaber, konturatuko da corpusei buruzko informazioa biltzeko saio batzuk egin direla, baina oraindik ere barreiatu samarra dagoela. Datu bila, zenbait erreferentzia-gunetara jo dugu lehenik (corpus-baliabideen zerrendak, katalogoak, banatzaileak...), gehienetan Internet bidez; bigarren, corpus batzuen argitalpenetara eta guneetara.

Hasieran euskarazkoak ez diren corpus eleaniztunak ere aintzat hartzea pentsatu genuen, baina dagoen produkzio-kantitate demasa ikusita, bestelako ahalegina eskatzen duela erabaki dugu, eta beste baterako utzi.

Datu-bilketaren emaitza artikulu honetan aurkez daitekeena baino handiagoa gertatu da. Datu-bilduma osoan interesa duenak Elhuyar Fundazioaren Hizkuntza Zerbitzuen guneko I+G+b azpisaileko 'Argitalpenak' ataletik eskura dezake "Corpus-baliabideen dossierra" dokumentua. Hurrengo ataletan, alderdi esanguratsuenak laburbilduko ditugu, eta gogoeta xume bat azaldu.

4.1 Oinarrizko erreferentzia orokorrak

Lehenik, ikuspegi orokorra lortze aldera, corpusen erreferentzia-gune hauetara jo dugu:

ELRA (European Language Resources Association)	http://catalog.elra.info/
LDC (Linguistic Data Consortium)	http://www.ldc.upenn.edu/Catalog/index.jsp
MULTEXT project	http://aune.lpl.univ-aix.fr/projects/multext/MUL4.html
TEI (Text Encoding Initiative)	http://www.tei-c.org/Activities/Projects/
Corpus survey	http://bowland-files.lancs.ac.uk/corplang/cbls/corpora.asp#_Toc92298891
Gateway to Corpus Linguistics	http://www.corpus-linguistics.de/html/corp/corpora_menu.html
David Lee-ren Bookmarks for Corpus-based Linguists	http://devoted.to/corpora
CLLT_OSU	http://cllt.osu.edu/lingCorpusLinks.html#corp
ICAME	http://khnt.hit.uib.no/icame/manuals/index.htm
Corpus Resources_Wirote Aroonmanakun	http://pioneer.chula.ac.th/~awirote/ling/corpuslst.htm

Hizkuntza guztiak kontuan hartuta, ingelesaren agerrera erabat gailentzen da besteen aldean, bai kantitatez bai corpus-moten aniztasunez. Bestalde, aitortu beharra dago euskarazko corpusen erreferentziarik ez dela izaten horrelakoetan (ELDAren bidez banatzen direnak alde batera utzita).

4.2 Corpusak

Aztertu dugun corpus-baliabide bakoitzaren ezaugarriak antolatzeko eta esparru erkagarriak eratzeko, aldagai hauek hartu dira kontuan: egilea, data, iturria (idatzia –argitalpen 'tradizionalak' nahiz Internet bidezkoak–, ahozkoa, edo biak), gaia (orokorra, espezializatua), asmoa (deskriptiboa, ereduzkoa), edukia, tamaina, mota (hizkuntza bat baino gehiagoko corpusetarako), hizkuntza, eremu geografikoa (dialektoak edo beste aldaeraren bat kontuan hartu diren), denbora (sinkronikoa, historikoa), prozesatzea (automatikoa, eskuzkoa edo bietatik), etiketatze-eredua/sistema, eskuragarritasuna, emaitzak, tresnak, salneurria, eta metadatuak nondik hartu ditugun (eskuragarritasunaren helbideko informazioa nahikoa ez bada).

4.2.1 Euskarazko corpusak

Euskarazko hamaika corpus identifikatu ditugu: batzuk euren buruari *corpus* esaten diote, beste batzuk corpus-izaera dute eta beste batzuk corpustzat hartzeko modukoak dira, corpus gisako erabilera bideratu dutelako (corpus-analisian erabili ohi den moduko kontsulta-aukerak eskaintzen dituztenak, esaterako).

Corpusa	Egilea	Data	Iturria	Mota	Asmoa	Tamaina	Etiketatzea
Orotariko Euskal Hiztegiaren testu-corpusa (OEHTC)	Euskaltzaindia	1984-2005	idatzizkoa	orokorra	deskriptiboa	6 M hitz	ez

XX. mendeko euskararen corpus estatistikoa (XXMECE)	Euskaltzaindia; UZEI	2002	idatzizkoa eta ahozkoa (transkribatuta)	orokorra, orekatua, lagindua	deskriptiboa	4,6 M hitz	automatikoa eta eskuz: SGML
Ereduzko Prosa gaur (EPG)	EHU eta Donostiako Udala	2007	idatzizkoa	orokorra	ereduzkoa	25,1 M hitz	lema eta ezaugarri morfologiko batzuk automatikoki
Zientzia eta Teknologiaren Corpusa (ZTC)	IXA Taldea (EHU) eta Elhuyar Fundazioa	2002-2008	idatzizkoa	berezia (zientzia eta teknologia); orekatua, lagindua	deskriptiboa	bertsio berrienerako: 8,5 M hitz; hortik 2 M inguru orekatuak	egiturazkoa eta linguistikoa: automatikoki eta eskuz. XML TEI-P4
Klasikoen Gordailua (KG)	Susa literatura-argitaletxea		idatzizkoa	berezia: euskal literatura klasikoa	deskriptiboa	11,9 M hitz	lematizatua?? TEI
Ibinagabeitia Proiektua (IP)	Susa literatura-argitaletxea	2000-2004	idatzizkoa	berezia: literatura-aldizkarien gordailua	deskriptiboa	451 aldizkari ale; 7.949 artikulu	linguistikoa: lematizazioa
FonAtari	Deustuko Unibertsitatea eta Bizkaiko Foru Aldundia	2001etik aurrerakoa	ahozkoa nagusiki, transkripzio eta iruzkinekin; bideoak ere bai	orokorra	deskriptiboa		linguistikoa: unitate foniko txikienak, hitzak, esaldiak eta testuak
Bizkaifon	Aholab (EHU) eta Bizkaiko Foru Aldundia	2003tik eskuragarri ELDAn	ahozkoa eta transkripzioak	orokorra	deskriptiboa	21 grabazio ordu: 11.569 hitz	SGML TEI-P4
Basque Spoken Corpus (BSC)	Jon Aske (Salem State College)	1993ko datuak; ELDAn 2002tik	ahozkoa, transkripzio eta deskripzioekin	berezia: bi film laburren azalpenak	deskriptiboa	42 narrazio eta 53 jardun libre	
Basque FDB-1060 database (BFDB)	Aholab (EHU)	ELDAn 2003tik	ahozkoa, dagokion transkripzioa eta deskripzioekin	berezia; ahozko jardun gidatua	deskriptiboa	45.580 item (1.060 lekuko)	transkripzioak SAMPA bidez; SpeechDat erako datu-basea
Euskararen Prozesamendurako Erreferentziazko Corpusa (EPEC)	IXA Taldea (EHU)	egiten	idatzizkoa			50.000 hitz; aurreikuspena: 200.000	morfologikoa eta sintaktikoa: automatikoa gehi eskuzkoa

Horiez gain, Eusko Jaurlaritzaren IKT inbentarioan[2] *Euskarazko Testu Corpusa* izeneko baliabidea ageri da, 25 milioi hitzekoa, eta 'Euskarazko ahotsa sintetizatzeko eta ezagutzeko motorrak' garatzeko bildu bide dena (uste dugu EJk Scansoft-i esleitutako *Aditu* aplikazioaren garapenean erabili dela). Bestetik, nahiz eta, esan dugunez, lan honetan corpus eleaniztunen arloa ez dugun jorratuko, ezin esan gabe utzi Interneten euskarazko bi corpus eleaniztun kontsultagai daudela: LEGE-bi corpusa (Deustuko Unibertsitatea) eta Eroski-ren *Consumer* aldizkariaren español-galego-catalán-euskara corpusa. Deigarria da, gainera, bi corpus horiek CLUVI-*Corpus Lingüístico da Universidade de Vigo* gunean egotea[3]. Lehenak 2,4 milioi hitz inguru ditu guztira, eta bigarrenak 5,6 milioi. Azkenik, EHUko Euskara Zerbitzuak argitaratzen dituen liburuen itzulpenen kontsulta eskaintzen du (jatorrizkoa eta euskarazko itzulpena)[4].

Badira *sensu stricto* 'corpus' ez diren testu-biltegiak, baina, linguistikoki prozesatuta daudenez edo bilatze-aukera bereziak eskaintzen dituztenez, interesgarriak izan daitezkeenak. Batzuk aipatzearren, eta dagoen aniztasuna erakuste aldera:

Armiarma; *Bonaparte Ondarea*; *Deba Ibarreko Ahotsak*; *EHU-Euskarazko Testu Biltegia*; *Eibar.org Materixala*; *Eibartarren ahotan-Eibarko Fonoteka*; *Euskal Herriko hizkuntz atlasa: ohiko euskal mintzamoldeen antologia*; *Euskal testuen gordailua*; *HABE Ikasbil*; *IVAP Itzulpen Memoriak*; *Teatro testuak*; *Zientzia.net*.

4.2.2 Beste hizkuntza batzuen egoera

Erdal corpusen inbentario exhaustiboa egitea gure ahalmenetik haraindi dago, eta, seguru aski, ez da nahitaezkoa corpus-baliabideen oinarrizko azterketa konparatiboa egiteko. Beraz, zenbait hizkuntzaren corpus 'nagusiak' arakatu eta aztertu ditugu, hirurogeita hamar guztira, hizkuntza hauetakoak: ingelesa, frantsesa, gaztelania, katalana, galegoa, alemana, arabiera, daniera, errusiera, eslovakiera, greko (modernoa), hungariera, italiera, irlandera, kroaziera, nederlandera, norvegiera, poloniera, portugesa, suediera, suomiera, txekiera.

4.3 Zenbait gogoeta

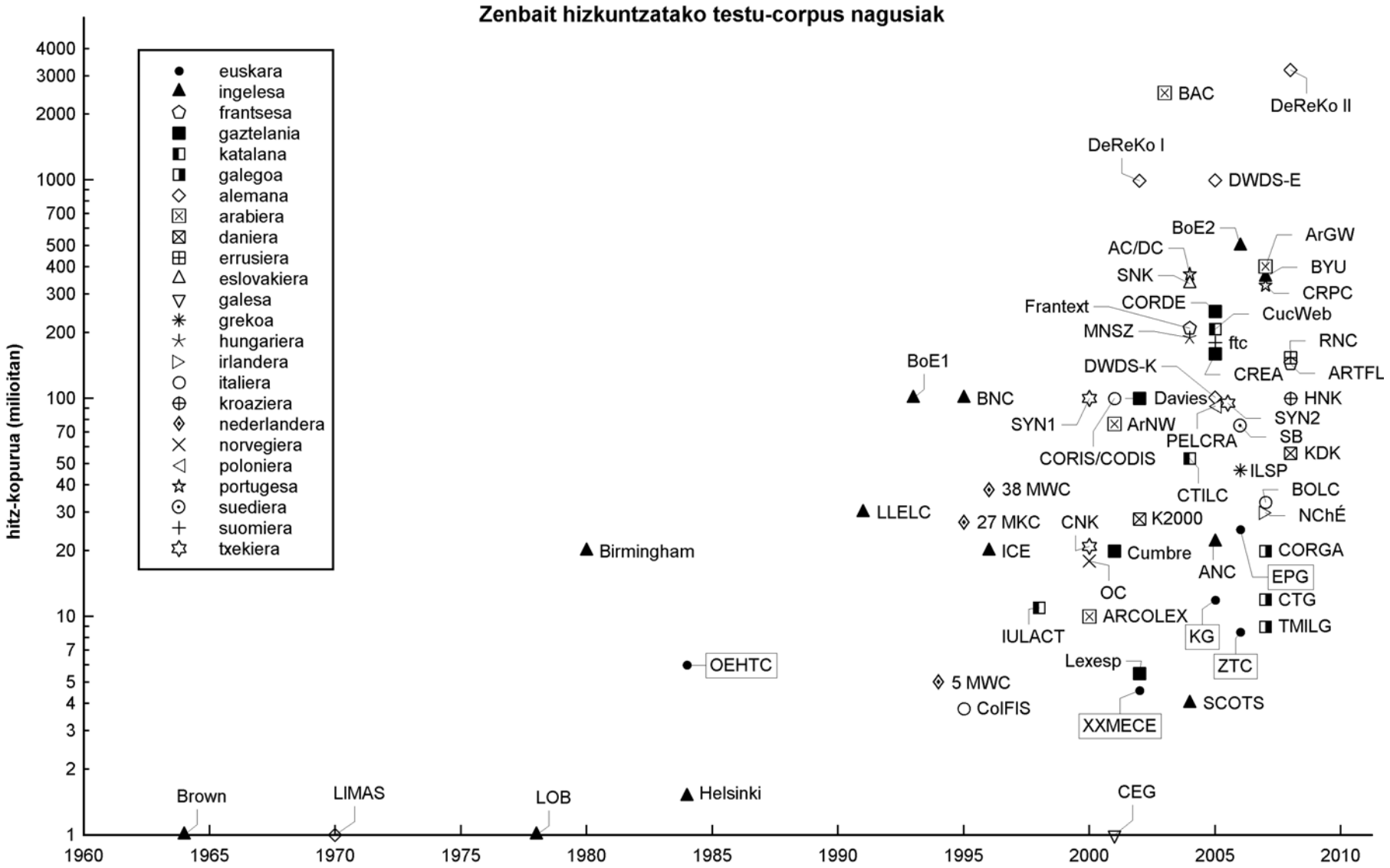
Corpus-tipologiari begiratzen badiogu, esan daiteke euskarazko produkzioak, bere txikian, gutxieneko aniztasuna agertzen duela. Erreferentzia-corpus diakronikotzat har litezkeen bi corpus dauzkagu, egungoak ez diren arren (OEHTC eta XXMECE), corpus berezi bat (ZTC; Alegria *et al*. 2005b, 2006b; Areta *et al*. 2007), literatura- eta prentsa-corpus handi bat (EPG, hein batean ere 'berezi' dena), euskal literatura-klasikoen eta -aldizkarien bilduma (KG), etiketatze sintaktikoa eta semantikoa duen corpus aurreratu bat (EPEC)… Nazioartean, corpus 'nagusi' gehienak orokorrak dira. Bereziak edo espezializatuak ere badira, noski, eta horien artean, corpus zientifiko-teknikoak dira ugarien orokorrean; euskaraz maila horretako urratsa eginda dagoela esan daiteke, behintzat.

Tipologia aldetik, kasu berezia da ahozko hizkuntzarena. Aipatu berri ditugun 'testu'-corpus horietan ez da ahozko hizkuntza jasotzen[5]. Hizketa-corpusak (*speech corpus*) ere baditugu, baina ezin dira aurrekoekin konparatu. Gehienak hizketaren teknologiak garatzeko eratuak dira, edo euskalkien ikerketarako pentsatuak. Beste hizkuntzetako testu-corpus esanguratsuetan, idatziak pisu handiagoa izaten du (prozesatzen errazagoa baita), baina ahozko laginak sartu dira corpusetan, gehiago edo gutxiago (BNCn, ahozkoa % 10 da; ICEn, % 60). Euskaraz *ahozko corpus* izenez (*spoken corpus*) bataiatutako proiektu bateratu eta erreferentzial baten hutsunea sumatzen dugu; gainera, gure ustez argi dago erreferentzia-corpus batek hizketa hein batean behintzat jaso behar lukeela.

Ingelesezko corpusgintzaren bilakaera jasotzen duen A. Renouf-ek egindako eskemarekin alderatu ahal izateko, aztertutako corpus 'nagusiak' kokatu ditugu hurrengo orrialdeko grafikoan. Hizketaren teknologiak garatzeko corpusak, *speech corpus* direlakoak, ez ditugu irudian sartu. Bestetik, corpusak amaituta dudenean, amaiera-urtea zehaztu dugu; corpus-proiektua amaitu gabe badago, edo etengabe elikatzen bada, azken eguneratze-lanen dataren

eta hitz-kopuruaren datuak ageri dira irudian. Kontuan hartu corpus-proiektu batzuk iraupen luzekoak direla; esaterako, gaztelaniazko CREA corpusa 1997an hasi zen, eta gaur egun ere elikatu egiten da.

Hurrengo orrialdeko grafikoan erabili diren corpusen akronimoen azalpena eranskinean eman dugu.


Zenbait hizkuntzatako testu-corpus nagusiak
hitz-kopurua (milioitan)
4000
3000
2000
1000
700
500
400
300
200
100
70
50
40
30
20
10
7
5
4
3
2
1
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
euskara
ingelesa
frantsesa
gaztelania
katalana
galegoa
alemana
arabiera
daniera
errusiera
eslovakiera
galesa
grekoa
hungariera
irlandera
italiera
kroaziera
nederlandera
norvegiera
poloniera
portugesa
suediera
suomiera
txekiera
BAC
DeReKo II
DeReKo I
DWDS-E
BoE2
ArGW
AC/DC
BYU
SNK
CRPC
Frantext
CORDE
CucWeb
MNSZ
ftc
RNC
BoE1
DWDS-K
CREA
ARTFL
BNC
Davies
HNK
SYN1
ArNW
SYN2
PELCRA
SB
CORIS/CODIS
KDK
ILSP
38 MWC
CTILC
BOLC
LLELC
27 MKC
CNK
K2000
NChÉ
ICE
Cumbre
CORGA
ANC
EPG
OC
CTG
ARCOLEX
IULACT
KG
TMILG
Lexesp
ZTC
5 MWC
SCOTS
ColFIS
XXMECE
Birmingham
OEHTC
Brown
LIMAS
LOB
Helsinki
CEG

Corpusen argitaratze-datei erreparatuz, nabarmentzekoa da euskarazko lehen corpusa, OEHTC), 1984an amaitu zela, eta handik gutxira hasi zela prestatzen EEBS (*Egungo Euskararen Bilketa-lan Sistematikoa*), gerora XXMECE izango zena. XXI. mendearen hasieran, ZTC, KG eta EPG argitaratu dira. Beste hizkuntzen kasuan, ezaguna da ingelesa aitzindaria izan dela, eta alemana nahikoa aurreratua izan da alde horretatik. Gainerako hizkuntzetan, gaur egun ezagunak diren corpus handi gehienak ere XX. mende amaierakoak eta XXI. mende hasierakoak dira. Beraz, gure iritzia da euskara ez zela 'berandu' iritsi corpusgintzara, ez behintzat beste hizkuntza nagusi asko baino askoz beranduago. OEHTC posizio aurreratuan ageri da grafikoan, eta argi dago corpus hori orduan egin izana ikuspegi estrategiko baten seinaletzat hartu behar genukeela (nahiz eta lematizatu gabea izateak nabarmen murrizten duen haren baliagarritasuna), baina aurrerago, XX. mendearen amaieran, moteltze nabarmen bat gertatu da, eta corpus 'handien' eraketan hamarkada baten atzerapena dugu, gutxienez, inguruko hizkuntzekiko.

Atzerapen hori nagusiki erreferentzia-corpusei dagokie, eta, nabarmen, tamainari. OEHTC eta XXMECE lorpen handiak izan dira, baina azpimarratzekoa da ez dagoela oraingoz euskarazko corpus 'erraldoirik', eta neurri 'txikiko' baliabideak direla aitortu behar da. *Ereduzko Prosa gaur* da, tamaina aldetik, nabarmenena, baina ezin genezake erreferentzia-corpustzat hartu (orekatu gabea da). Nazioartean ere antzeko zerbait sumatu dugu: zenbat eta 'txikiagoa' hizkuntza (hedaduraz, hiztunez, diruz/aurrekontuz...), corpus apalagoak egiten dira; zenbat eta 'handiagoa' (hizkuntza), hainbat eta ugariagoak emaitzak, alde guztietatik begiratuta. Aipagarria da alemanezko corpus batzuen neurri eskerga (*DWDS Ergänzungscorpus* eta DeReKo), adibidez, baina hizkuntza horren corpusgintza tradizio helduak azalduko luke hori hein batean. Nabarmentzekoa da, bestalde, hain 'handi' ez diren hizkuntza batzuetan oso corpus handiak eratu direla: eslovakiera (SNK), hungariera (MNSZ), suomiera (ftc), txekiera (SYN2000)…

Corpusen egituratze- eta prozesatze linguistikoa dela eta, euskarazko corpus batzuek betetzen dituzte gaur egungo estandarrak. Adierazi dugu OEHTCn informazio linguistikoa (lema, kategoria...) ez izatea tamalgarria dela, baina geroztik egin diren testu-corpus nagusietan behintzat ez da hutsegite hori berriz gertatu (XXMECE, ZTC, EPG...). Bestetik, XXMECE eta ZTC corpusak, bederen, corpusak etiketatzeko nazioarteko estandarren arabera landu dira (hurrenez hurren, CES eta TEI P4). Etiketatze-lan horiek egiteko, tresna automatikoak erabiltzen dira gaur egun; 6. atalean jardungo dugu gai horretaz.

Egileen alorrean, zenbait erakunde-motak bultzatu dute euskal corpusgintza: erakunde akademikoak (hizkuntzaren akademia, bertako unibertsitate bi eta atzerriko beste bi), hizkuntza-zerbitzuak eskaintzen dituzten eta ikerketa-atala duten enpresak, argitaletxe bat, nazioarteko erakunde bat[6], eta, azkenik, norbanako bat[7]. Ekimen batzuk (OEHTC, XXMECE, EPG...) diru publikoz osorik finantzatu dira, eta gainerakoek diru-laguntza publikoa jaso dute. Beste hizkuntzen kasuan ere, nabarmena da unibertsitate eta entitate akademikoen e(ra)gile- eta garatzaile-lana, instituzioen eta, hainbatetan, enpresa handien finantza-laguntzaz. Sarritan (mundu anglosaxoian batez ere), proiektu komertzial handiek ekarri dute corpus-proiektu

handiak abian jartzea (*The Bank of English-Cobuild*, esaterako), batez ere baliabide linguistikoak eratzeko, nahiz eta oinarrizko baliabidea bera egitea ere helburu nagusi izan den ‘corpus-proiektu nazionalen’ kasuan: ohikoak dira *erreferentzia-corpus*, *corpus nazional* izendapenak. Hizkuntza asko hartzen dituzten makroproiektuak ere aipatzekoak dira: LDCren *Gigaword*, *Wacky* proiektua, aipatu berri dugun *SpeechDat*, *C-ORAL-ROM* eta abar. Banakoen corpus handiak ere badaude nazioartean (aipagarriena, arabierazko T. Buckwalter-en corpus erraldoia, kasik hiru mila milioi hitzekoa). Oraingoz behinik behin, ez da horrelakorik.

Aipatzekoa ere bada euskarazko corpusen asmoa deskriptiboa dela, batean izan ezik (EPG). Hizkuntza-ereduari dagokionez, berrienak euskara batuari so daude (tresna automatikoak horrekin dabiltza hobekien), baina euskalkien eta beste barietate batzuen presentzia ere badago. Beste hizkuntzetako corpusen diseinuan ere, corpusa 'adierazgarria' eta 'orekatua' lortzea da irizpide nagusia, dela hizkuntza oro har aztertzeko (erreferentzia-corpusetan) dela alderdi edo erabilera-arlo jakin bat aztertzeko (corpus berezietan). Horrela ulertuta, baliabide deskriptiboen erabateko gailentasuna dago. Horrelakoetan, aldagai estralinguistikoei erreparatzen zaie sailkapenak, geruzak eta abar antolatzerakoan (genero, eremu, egileen datu soziolinguistikoak...). Hizkera-aldaerak (estandarraz gain) oso presente daude mundu zabaleko corpusetan, eta hori koherentea da aipatu berri den deskriptibismoarekin.

Eskuragarritasunak aipamen berezia merezi du: euskarazkoetan, Interneteko doako kontsulta da testu-corpusen banatze-bide ohikoena. *Klasikoen Gordailua*-ren gunetik corpus gordina (linguistikoki prozesatu gabea) deskarga daiteke. Bestetik, IXA taldeak eta Elhuyar Fundazioak iragarri dute ZTC ELDAren katalogoan jarriko dela 2008an[8]. Hizketa-corpusen kasuan, berriz, hiru baliabide daude ELDAren katalogoan (Bizkaifon, BSC eta BFDB).

Corpusaren kontsulta-sistema Interneten jartzea lehen urrats bat da, ezinbestekoa, baina jakina da erabilera askotarako ez dela aski (hizkuntza-ikerketa, baliabide lexikoak eratzea, hizkuntza teknologien garapena...). Beste hizkuntzetan ere denetik dago, baina corpusen eskuragarritasuna handiagoa da oro har, arrunki aipatzen diren copyright-arazoak gorabehera.

5 Corpusak eta Internet

Aurreko atalean aztergai izan ditugun baliabideak corpus izateko asmoz diseinatutakoak eta eratuak dira (gehienak behintzat, badira testu-biltegi direnak baina corpus gisa ere erabil litezkeenak, edo erabili izan direnak). A. Renouf-en eskeman, ordea, corpus-metodologia berria agertzen da 1998tik aurrera, eskeman ‘The Web as corpus’ eran aurkeztua, baina oro har berak *cyber-corpus* terminoaz adierazten duena.

Bi eratako ikuspegi daude Internet eta corpusei dagokienean. Lehenengoa Internet zuzenean corpus bat balitz bezala kontsultatzea da (*web as corpus*). Bide horretatik, hainbat tresna egin dira Interneten hitz bat edo batzuk bilatzeko aukera ematen digutenak, bilatzaileek egiten duten gisara, baina emaitza, dokumentu-zerrenda izan beharrean, dokumentu horietako

agerpenak euren testuinguruetan erakusten dituztenak. Honelako tresnen adibide dira *WebCorp*[9] (Kehoe & Renouf, 2002), *WebCONC*[10] edo *KWiCFinder*[11] (Fletcher 2001).

Bigarren ikuspegia da Internet corpusa eratzeko edo elikatzeko testu-iturritzat erabiltzea (*web for corpus*). Hori bi modutara egin ohi da: orri edo esteka jakin batzuetatik abiatuta estekak jarraitzea, edo hitz batzuetatik abiatuta bilatzaileak erabiltzea. 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 *ItWaC* eta *DeWaC* italierazko eta alemanezko corpusak, 2 mila milioi eta 1,7 mila milioi hitzekoak, hurrenez hurren. Gainera, *Corpus building for minority languages* gunean[12], 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).

Euskarak ere heldu dio Internet eta corpusen gaiari. Elhuyar Fundazioaren bi proiektu dira aipagarriak hemen, azaldu ditugun bi ikuspegietan oinarrituak. Lehena *CorpEus* proiektua da, Internet euskarazko corpus erraldoi gisa baliatzea helburu duena (Leturia *et al*., 2007a). *WebCorp* eta horien antzeko zerbitzu bat ezarri da[13], baina euskararen berezitasunei egokitua. Bigarrena *Co3* proiektua da (*Comparable Corpus Compiler*), Internetetik corpusak osatzeko testuak automatikoki eratuko dituen tresna. Batez ere corpus eleaniztun konparagarriak lortzera bideratuta badago ere, euskarazko corpus elebakarrak egiteko ere balioko du, bai corpus orokor handiak bai espezializatu txikiagoak[14].

Aipatzeko modukoa da *CorpEus* eta *Co3* proiektuetan hizkuntza-teknologien beste alor batzuetarako (informazio-bilaketarako, adibidez) oso baliagarriak diren aurrerapenak lortu direla. Bi proiektu horiek Interneteko bilatzaileak erabili behar dituzte ezinbestean, eta ezagunak dira bilatzaileak euskararako ez direla oso egokiak. Batetik, morfologia aberatseko beste hizkuntzetan bezala, ezinbestekoa da bilaketa lema bidez egitea emaitza onak lortzeko, eta hori ez dute egiten. Bestetik, ezin zaizkie euskarazko emaitzak soilik eskatu (ezta beste hizkuntza txiki gehienenak ere). Bi arazo horiek konpontzeko, teknika bana garatu da proiektu horietan: bilagaia sorkuntza morfologikoaren bidez hedatzea eta hitz bereizgarrien bidezko hizkuntza-iragazkiak. Teknika horien bidez garatu da, adibidez, *Elebila* Interneteko lehen euskarazko bilatzailea (Leturia *et al*., 2007b).

A. Renouf-en eskemari bagagozkio, esan genezake arlo honetan euskara aski aurreratua edo ‘egunean’ dagoela.

6 Corpus-tresnak

Ez da lan honen helburua corpusak egiteko eta ustiatzeko behar diren tresnetan sakontzea, baina ezinbestekoa, gaingiroki bada ere, oinarrizko ideia batzuk agertzea.

Corpusak egitea lan handia eta konplexua da, eta gaur egun tresna informatikoak ezinbestekoak dira corpusgintzan. Bestetik, corpusak kontsultatzeko, ustiatzeko eta haietatik

informazioa lortzeko ere ezinbestekoa da tresna automatikoak izatea. Corpus-tresnen azalpena hiru ataletan banatuko dugu.

6.1 Corpusak egiteko eta kudeatzeko tresnak

Corpusa diseinatu ondoren, lehen urratsak dira corpuseratuko diren testuak hautatzea, biltzea, corpuserako hautatutako formatu bateratu batera bihurtzea, eta informazio bibliografikoa (metadatuak) eta egiturazkoa etiketatzea. Diseinuaren eta aipatu urrats horien ezaugarrien arabera, corpus-proiektuak asko aldatzen dira[15]. Oso zaila da corpus guztien ezaugarrietara egokitzen den tresna estandarrak egitea. Horregatik, corpus-proiektu gehienek bere kudeatze-tresna garatu dute, edo ezer berezirik prestatu gabe moldatu dira bestela.

Euskarari dagokionez, corpusak eratzeko eta kudeatzeko erabilitako tresnen inguruan argitaratuta dagoen informazio bakarra (guk dakigula behintzat) *Corpusgile*-ri buruzkoa da (Alegria *et al*., 2005). ZTC eraikitzeko erabilitako tresna da. Corpus orekatu bat egiteko urrats guztiak hartzen ditu bere baitan (inbentarioa, laginketa, bihurketa eta egitura-etiketatzea) eta etiketatze linguistikoko tresnak ere bertan integratzen dira (nahiz eta aparteko aplikazioak izan).

6.2 Etiketatze linguistikorako tresnak

Ahal dela, corpusak linguistikoki prozesatu behar dira, gero corpusetik informazio linguistikoa lortu ahal izateko. Prozesatze horren bidez, testuko formen lema, kategoria, kasua, funtzio sintaktikoa, adiera eta abar etiketa daitezke. Etiketatzea txertatua (dokumentuan bertan) edo banatua (dokumentuaz aparteko fitxategi batean) izan daiteke.

Gaur egun, lan hori automatikoki egiten da normalean, LNPko tresnen bidez, eta gero, batzuetan, eskuz berrikusi edo desanbiguatzen dira analisi automatikoak. Tresna horiek hizkuntzaren araberakoak izaten dira, baina badira hizkuntzatik independenteak direnak ere, hizkuntzaren lexikoia eta eskuz etiketatutako testu-multzoa edukita betiere (*TreeTagger*, adibidez). Bestetik, ez dira berariaz corpusak egiteko garatutako tresnak, beste erabilera anitz dituzten tresna linguistikoak baizik.

Euskarazko corpusetan zenbait tresna erabili dira. ASP enpresaren *Kapsula* izeneko tresna da horietako bat. Testuetako hitzen lemak ateratzen ditu, eta gero hitzen lemak indexatu eta horien bidezko bilaketak egiteko aukera ere ematen du. EPG, IP eta KG prozesatzeko erabili da.

XXMECEren parte bat (1900-1991) Hizkia enpresaren *RTerm* tresnaren laguntzaz lematizatu zen (eskuzko lematizazio batetik abiatuta); gainerakorako (1991-1999), IXA taldearen eta UZEIren *Euslem* tresna erabili zen, eta gero eskuz berrikusi eta desanbiguatu ziren hitz guztiak (Urkia, 2002).

ZTC linguistikoki etiketatzeko, *Euslem*-en garapen berria den IXA taldearen *Eustagger* tresna erabili da. Sintaxi-mailaraino etiketatzen du, TEI formatuan, anotazio banatua erabiliz. Gero, zati bat (corpusaren gune orekatua) eskuz berrikusi eta desanbiguatu da lema- eta kategoria-mailan, EULIA tresnaren bidez (Artola *et al.* 2004). *Euslem* eta *Eustagger*

etiketatzaileek datu-base lexikal bat erabiltzen dute (EDBL-Euskararen Datu Base Lexikala; Aldezabal *et al*. 2001), eta erabiltzaileak sortutako lexikoi osagarria ere erabil dezakete.

Berriki, *EnergiaTech* enpresak *BasqueLem* lematizatzailea iragarri du.

6.3 Corpusak analizatzeko eta ustiatzeko tresnak

Corpusa bera amaituta dagoenean, bertatik informazioa lortzeko tresnak behar dira. Bilatzeko aukeren eta ematen den informazio-motaren arabera, zenbait tresna-mota edo 'belaunaldi' daude: a) agerpenak eta testuinguruak erakusten dituztenak, eskuarki KWIC (*key word in context*) edo 'konkordantziak' izeneko formatuan; b) agerpenen eta testuinguruko agerpenen (kolokazioen) estatistikak ere ematen dituztenak; eta c) bilagaien emaitzak gramatika-ezaugarrien, konbinazio lexikoen edo semantika-erlazioen arabera antolatuta ere eskaintzen dituztenak (Kilgarriff *et al*. 2004).

Euskarazko corpusak kontsultatzeko ia tresna guztiak lehen motakoak dira, hau da, gehienez ere bilagaiaren testuinguruak eta maiztasunak erakusten dituzte. ZTC da bigarren motakotzat jo daitekeen bakarra, hitzaren aurreko eta ondorengo testuinguruan agertzen diren formak/lemak eta horien maiztasunak erakusten baititu, taulatan zein grafikotan.

Corpusak kontsultatzeko tresnetan badaude erabilera zabaleko tresnak (nahiz eta norberak bere corpusarentzat berariaz egindakoak ere asko diren). Corpusak kontsultarako prozesatzen dituzten tresna ezagun batzuk: *IMS Corpus WorkBench*[16] (corpuseko dokumentuak kudeatzeko atala ere baduena), BNCren *Xaira*[17], DWDSren DDC[18], ARTFL-Frantext corpusaren *Philologic*[19] eta *Sketch Engine*[20] (azken hori hirugarren motakoa da). Horiez gain, konkordantzia-aplikazioak ere badaude (WordSmith Tools, Concordance...).

Corpusak kontsultatzeko tresnez gain, corpusetatik informazioa automatikoki erauzteko tresnak ere oso baliagarriak dira, terminoak eta kolokazioak erauztekoak adibidez. Euskararako badaude Elhuyar Fundazioak eta IXA Taldeak egindako horrelako tresnak: *Erauzterm*, corpus elebakarretatik terminoak erauzteko tresna (Alegria *et al*., 2005a), eta *ELexBI*, itzulpen-memorietatik termino-bikote elebidunak erauztekoa (Alegria *et al*., 2006a).

7 Etorkizunari begira

7.1 Erreferentzia-corpusa

Erdal corpusen azterketak agerian utzi duen alderdi bat da hizkuntza askotan eratu direla erreferentzia-corpusak izateko asmoz diseinatutako proiektuak. Batzuek termino hori bera daramate izendapenean, eta beste batzuek, beharbada BNCren izendapenean inspiratuta, 'corpus nazional'.

Erreferentzia-corpusen garrantziaz ohartarazi gaituzte hainbat adituk (Leech 2002), eta, beharbada, alderdi hori da euskarazko corpusetan dagoen gabezia nabarmenena. Azken urteetan, maiz hitz egiten da euskararen erreferentzia-corpusaren proiektua egiteko dagoen premiaz. Euskaltzaindiak eta Eusko Jaurlaritzak berak ere aipatu dute beren planetan, eta

badirudi garaia heldu dela horrelako proiektu bati ekiteko. Aurreproiektu-proposamen bat ere badago, M. Urkiak landua eta aurkeztua hain zuzen ere (Urkia 2005).

Gure asmoa hemen ez da nolako erreferentzia-corpusa behar genukeen proposatzea, baina bai behintzat erreferentzia-corpusa bat diseinatzean kontuan hartu behar diren zenbait alderdi, erantzun behar diren galdera batzuk eta mahai gainean argi jarri behar diren erabakigaiak agertzea:

- Zer dugu gogoan *erreferentzia-corpus* terminoa erabiltzen dugunean?
  - Hizkuntzaren lagin adierazgarria izateko diseinatu den corpusa?
  - Hiztunei 'ereduzkotzat' eskaintzen zaizkien testuez osatutako corpusa?

  Gure iritzia da lehen ideia dela *reference corpus* terminoaren jatorrizko adierari dagokiona[21]; baina horrelakoa da gaur egun euskaldunok behar dugun corpusa? Hizkuntza 'normalizatuetarako' egindako definizioak balio du bere horretan normalizazio bidean den hizkuntza minorizatu baterako?

  Zertarako nahi dugu? Gaur egungo euskara aztertzeko? Gaur egungo euskaldunek erabiltzen duten euskara aztertzeko? Gaur egun irizpide batzuen arabera kalitatekotzat jotzen diren euskarazko testuak aztertzeko, eta horietako euskara ereduzkotzat proposatzeko?
- Era batera edo bestera, adierazgarritasunaren eta orekaren planteamendu bat egin behar da (Biber 1993). Zeren adierazgarri? Nola lortuko dugu populazioaren lagin adierazgarri bat, edo nola hurbil gaitezke? Zein da populazioa? Osorik ezagutu edo katalogatu behar da? Zein parametro erabiliko da obrak sailkatzeko? (gero laginketan eta orekatzean erabiltzeko). Nolako lagintze-eredua? (proportzionala, geruzatua...). Nolako laginak? (Obra osoak ala obra-zatiak? Obra-zatiak badira, zenbat obra bakoitzetik? Nola lagindua, jarraitua, ala lagin etenak? Zenbat hitzekoak?)
- Corpus irekia? (une oro adierazgarri izaten jarraitzeko mantendua)
  - Hasiera-urtea: XXI. mendearen hasieratik aurrera, ala zertxobait atzeragotik hasita, Euskaltzaindiaren ‘araugintza berritik’ (1990)?
  - Nola lortu corpusa etengabe elikatzea eta adierazgarritasunari edo orekari eutsi?
- Tamaina: zenbat hitz behar dira gutxienez gaur egungo euskararen lagin adierazgarria edo orekatua osatzeko? Erreferentzia-corpus askoren helburua den 100 milioi hitzeko tamaina lor liteke euskaraz?
- Testu-bilketa: zein lan egin behar da testu-hornitzaileekin (argitaletxeekin, egileekin...) corpus-proiektuaren beharraz eta haiek egin dezaketen ekarriaz ohartarazteko?

- Ahozkoaren tokia: corpusetik zenbat izango da ahozko hizkuntzatik lagindua? Zein erregistro eta diskurtso-mota corpuseratuko da? (lagunartekoak, formalak...; hizketa librea/kontrolatua...)
- Euskalkiak eta bestelako aldaerak: etiketatzaile automatikoak euskara estandarrerako daude optimizatuta; euskalkian idatzitako testuak corpuseratuko dira? Nola prozesatuko dira, eskuz? Etiketatzaileak euskalkietarako garatu edo moldatuko dira?
- Informazio linguistikoa (prozesatze linguistikoa): zein informazio etiketatuko da automatikoki (lema, kategoria, azpikategoria, kasua, funtzio sintaktikoa, adiera...) eta zein berrikusi, zuzendu eta desanbiguatuko da eskuz?

Urtetan geldi egon ondoren, badirudi laster etorriko dela M. Urkiaren proposamenei eta egin berri ditugun galderei erantzuteko garaia. Sagarnak behintzat horrela iragarri du Euskaltzaindian sartzeko hitzaldian:

"Euskaltzaindiak badu dagoeneko bide hori urratzen joateko egitasmoaren lehen zirriborroa. Bertan corpus monitore bat eraikitzen joatea planteatzen da, alde batetik euskara idatzizko hedabideetan ia unean-unean nola erabiltzen ari den jakiten lagunduko duena eta bestetik gorputza hartzen joango dena erreferentzia corpuseranzko bidean."

Hala izango ahal da!

7.2 Internet eta corpusak

Web-aren sorrerak aukera berriak eskaini dizkio corpusgintzari eta corpus-hizkuntzalaritzari, baina orain arte ez bezalako erantzunak eskatzen dituzten galdera berriak ere eginarazi dizkigu. Aukerak bistakoak dira: etengabe elikatzen den testu-bilduma erraldoia, dagoeneko digitalizatuta, erabiltzaile guztiek kontsultatzeko moduan. Corpus handiak biltzea kostu handiko lana izaten da, eta corpus monitore bat mantentzea ere eskakizun handikoa da; horra, bada, hizkuntzaren bilakaera datu handiak erabiliz eta arrazoizko kostuan aztertzeko baliabidea. Baina, hizkuntzaren azterketaren ikuspegitik, zenbait galdera eta eztabaidagai jarri ditu aukera horrek mahai gainean: datuen ‘ezegonkortasuna’ edo ‘errepikaezintasuna’, hizkuntza-estilo eta erregistro berezien presentzia (ia web-ean soilik agertu ohi direnak), hizkuntzaren 'kalitatearen' auzia (testu inprimatuetan ez bezala, Interneten orraztu gabeko testu asko dago, erregistro informalean idatzia eta abar...)... Adierazgarritasuna ere eztabaidagai da. Corpusaren analisi kuantitatiboak eta estatistikoak ere ezin dira 'corpus mugatuetan' bezala aplikatu edo interpretatu (batik bat, *off line* corpus batean ez bezala, 'populazioa' ezaguna ez delako) (Renouf 2007).

Dena den, web-a corpusaren definiziorik zorrotzenei lotzen ez bazaie ere, praktikan corpustzat hartzen dugunaren ezaugarriak betetzen ditu, eta orokorrean corpustzat edo corpus-iturritzat erabil daitekeela onartuta dago (Kilgarriff & Grefenstette 2003), eta horretan ari dira buru-belarri beste hizkuntzak. Kontuan izanda corpusetan (eta baita corpusgintzarako baliabideetan, tresnetan nahiz giza baliabide edo ekonomikoetan ere) euskara beste hizkuntza

askoren atzetik doala, beharrezkoa da Interneterako joera horretan sartzea, corpus 'kontrolatuak' (ohiko corpusak, erreferentziazkoa, corpus espezializatuak...) alde batera utzi gabe, noski.

7.3 Eskuragarritasuna eta zabalkundea

Arlo honetan euskarak zer hobetu nabaria duela uste dugu. Corpus gehienak Internet bidez kontsulta daitezke, baina esan dugu helburu askotarako ez dela aski. Gainera, kontuan hartu behar da corpus batzuk diru publikoz osorik finantzatuak izan direla, eta ez dela erraz ulertzen corpus horiek ez askatzea (ikerkuntzarako, esaterako) edo lizentzia baten truke ustiapen komertzialerako eskuragarri ez jartzea. Horrek ez du euskarazko corpusen erabilgarritasuna eta emankortasuna murriztu baizik egiten.

Bestetik, gure azterketan nabaritu dugu mundu-mailako ikusmiran euskarazko corpusak ez direla nahikoa ‘ikusten’, hau da, corpusen eta hizkuntza-teknologien erreferentzia-gune ezagunetan oso informazio gutxi aurkitu dugu euskarazko baliabideez. Arazo hori ez da euskarazko corpusena bakarrik, noski; esaterako, P. Bilbaok salatu duenez, Europako hizkuntza politika ‘txarrak’, edo ezegokiak, Europa mailan “ikusgabe bihurtzen gaitu” (Bilbao 2006). Eta hizkuntza eta bere baliabide eta produktuak ikusgai izatea, bestalde, osasun soziolinguistikoaren adierazletzat hartzen da, bizitasun/bizindar linguistikoaren adierazletzat (Corvalán 2005).

Nolanahi ere, nabarmentzekoa da ikerketa-talde eta erakunde batzuek egiten duten ahalegina euskarazko corpusen eta, oro har, hizkuntza-teknologien informazioa argitalpen eta biltzar espezializatuetan agertzeko. Horien denen bilduma luzeegia litzateke, baina artikulu honetan bildu dugun bibliografiak agerian jar dezake errealitate hori. Txanponaren beste aldea, berriz, corpus-proiektu batzuen inguruan dagoen erabateko dokumentaziorik eza da.

7.4 Estrategia bateraturik?

Euskararen munduan, hitzetik hortzera entzuten dugu ahaleginak batu, bateratu edo koordinatu egin behar direla, eta lankidetzaren beharra askok nabarmendu dute. Corpusen arloan, bistakoa da proiektu batzuetan erakunde publikoen bultzada ezinbestekoa dela, batik bat corpus handiak, erreferentzia-corpusak nagusiki, egingo badira, kostu handiko proiektu etengabe elikatu beharrekoak baitira. Horretan instituzio publikoek duten egitekoa argia da.

Bestetik, euskarazko corpusgintzak duen tamainarako, horretan ari direnen kopurua ez da txikia, eta gure ustez aniztasun hori ona da. Eragileen arteko lankidetza handiagoa eta proiektu bateratu gehiago sortzea litzateke, seguru aski, euskarazko corpusgintza indartzeko bideetako bat. Ikuspegi- eta tresna-aniztasuna ez daitezela oztopo izan corpusgintzan ditugun premiei behar bezala ez erantzuteko. Hemendik urte batzuetara, artikulu honetan aurkeztu dugun grafikoa eguneratzen dugunean, euskarazko corpus gehiago eta handiagoak ikusi nahi genituzke, eta, horien artean, euskararen erreferentzia-corpusa. Badugu zeregina.

8 Bibliografia

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G. & Lersundi M. (2001). "EDBL: A General Lexical Basis for the Automatic Processing of Basque." In *IRCS Workshop on linguistic databases. Philadelphia (USA).* [on line] [kontsulta: 08-03-31] <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1011897592/publikoak/2001-IRCS.pdf>

Alegria, I., Gurrutxaga, A., Saralegi, X. & Ugartetxea, S. (2005a). "Eraузterm: euskarazko terminoak erauzteko tresna erdiautomatikoa." In *Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera zientifiko-teknikoa*. Bilbo. [on line] [kontsulta: 07-05-28]. <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1141404023/publikoak/pdf>

Alegria, I., Areta, N., Artola, X., Díaz De Ilarraza, A., Ezeiza, N., Gurrutxaga, A., Leturia, I., Saiz, R., Sologaistoa, A., Soroa, A. & Valverde, A. (2005b). "Zientzia eta teknologiaren corpusa." In *Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera zientifiko-teknikoa*. Bilbo. [on line] [kontsulta: 07-05-28] <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1113384045/publikoak/ZT_Corpusa_Mendebalde.pdf>

Alegria I., Gurrutxaga A., Saralegi X., Ugartetxea S. (2006a). "ELexBI, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora." In *12th EURALEX International* Congress. pp 159-165 [on line] [kontsulta: 07-05-28] <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1177085533/publikoak/pdf>

Alegria, I., Areta, N., Artola, X., Díaz De Ilarraza, A., Ezeiza, N., Gurrutxaga, A., Leturia, I., Saiz, R., Sologaistoa, A., Soroa, A. & Valverde, A. (2006b). "Structure, Annotation And Tools In The Basque ZT Corpus." In *LREC-2006 5th International Conference On Language Resources And Evaluation.* Genoa. [on line] [kontsulta: 07-05-28]. <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1141404023/publikoak/pdf>

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 *Corpus Linguistics 2007*. Birmingham. [on line] [kontsulta: 08-03-31] <http://corpus.bham.ac.uk/corplingproceedings07/paper/65_Paper.pdf>

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sologaistoa A. & Soroa A. (2004). "EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora." In *LREC-2005 4th International Conference On Language Resources And Evaluation..* Lisboa [on line] [kontsulta: 08-03-31] <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1088448358/publikoak/04LREC_EULIA.pdf>

Baroni, M.& Bernardini, S. (2004). "BootCaT: Bootstrapping corpora and terms from the web." In *Proceedings of LREC 2004.* Lisbon, Portugal: ELDA, pp. 1313--1316. [on line] [kontsulta: 08-03-31] <http://sslmit.unibo.it/~baroni/bootcat.html>

Biber, D.. (1993). "Representativeness in Corpus Design." In *Literary & Linguistic Computing* 8. 243-257. orr.

Bilbao, P. (2006). “Europako Itun Konstituzionalak eskubide gabe utzi ditu hizkuntza-komunitate ugari”, *Uztaro*, 58

Christ, O. (1994). *"*A modular and flexible architecture for an integrated corpus query system." In *COMPLEX'94, Budapest*

Corvalán, G. (2005). “La vitalidad de la lengua guaraní en el tercer milenio en Paraguay”, <http://www.datamex.com.py/guarani/opambae_rei/tembihai/corvalan_vitalidad_del_guarani.ht>ml [2008-03-14]

Fletcher, W. H. (2001). "Concordancing the Web with KWiCFinder." *Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA, 23-25 March 2001. [on line] [kontsulta: 08-03-31] <http://webascorpus.org/searchwac.html>

Ide N. & Véronis J. (1995). *Text Encoding Initiative: Background and Context*. Kluwer Academic: Dordrecht.

Ide, N. (1998). "Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora." In *First International Language Resources and Evaluation Conference*, Granada, Spain.

Ixa Taldea & Elhuyar Fundazioa (2007) “Testu-corpusak: ezaugarriak, eraketa eta tresnak.” In *Hizkuntza komunikazioaren eta teknologiaren garaian*. Herri Arduralaritzaren Euskal Erakundea (IVAP). Gasteiz.

Kehoe, A. & Renouf, A. (2002). "WebCorp: Applying the Web to linguistics and linguistics to the Web." in *Proceedings of WWW2002*, Honolulu, Hawaii. <http://www.webcorp.org.uk/>

Kilgarriff, A.; Rychly, P.; Smrz, P.; & Tugwell, D. (2004). The Sketch Engine. In *Proc. of EURALEX 2004*, 105–116. [on line] [kontsulta: 08-03-31]<http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw>

Leech, G. (2002). "The Importance of Reference Corpora." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI. [on line] [kontsulta: 08-03-31] <http://www.uzei.org/corpusajardunaldia/06_gleech.pdf>

Leturia, I., Gurrutxaga, A., Alegria, I.& Ezeiza, A. (2007a). "CorpEus, a ‘web as corpus’ tool designed for the agglutinative nature of Basque." In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain, pp. 69--81. [on line] [kontsulta: 08-03-31} <http://www.corpeus.org/CorpEus%20WAC3.pdf>

Leturia, I., Gurrutxaga, A., Areta, A., Alegria, I. & Ezeiza, A. (2007b). "EusBila, a search service designed for the agglutinative nature of Basque." In *Proceedings of Improving non-English web searching (iNEWS’07) workshop*. Amsterdam, The Netherlands: SIGIR, pp. 47--54. [on line] [kontsulta: 08-03-31] <http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1190627800/publikoak/pdf>

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.

Rojo, G. (2002). "Sobre la lingüística basada en el análisis del corpus" In *Hizkuntza-corpusak. Oraina eta geroa. Donostia*: UZEI. [on line] [kontsulta: 08-03-31] <http://www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos54A.pdf>

Sagarna, A. (2007). "Euskara eta informazioaren teknologiak. Egungo egoeratik etorkizunera begira." *Euskaltzaindia - Sarrera-hitzaldia.* [on line] [kontsulta: 08-03-31] <http://www.euskaltzaindia.net/plazaberri/0024/gehigarriak/sagarna.pdf>

Scannell, K. P. (2007). "The Crúbadán Project: Corpus building for under-resourced languages." In *Building and exploring web corpora, Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain, pp. 5--15. [on line] [kontsulta: 08-03-31] <http://borel.slu.edu/crubadan/>

Sinclair, J, (1996). *Preliminary Recommendations on Corpus Typology*. EAGLES. [on line] [kontsulta: 08-03-31] <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>

Svartvik, J. (2007). "Corpus linguistics 25+ years on." In Facchineti, R. (Ed.) *Corpus linguistics 25+ years on*. Rodopi: Amsterdam- New York.

Urkia, M. (2002). "XX. mendeko euskararen corpusa." In *Hizkuntza-corpusak. Oraina eta geroa*. Donostia: UZEI. [on line] [kontsulta: 08-03-31] <http://www.uzei.org/corpusajardunaldia/03_murkia.pdf>

Urkia, M. (2007). "XX. mendeko euskararen corpus estatitistikotik XXI: mendeko erreferentzia corpusera." In *Espezialitateko hizkerak eta terminologia II. Euskara estandarra eta espezialitate hizkerak*. EHU: Leioa.

9 Eranskina

'Zenbait hizkuntzatako testu-corpus nagusiak' grafikoan erabilitako corpus-akronimoen azalpena:

Akronimoa	Corpusa	Hizkuntza
27 MKC	27 Miljoen Woorden Krantencorpus	nl
38 MWC	38 Miljoen Woorden Corpus	nl
5 MWC	5 Miljoen Woorden Corpus	nl
AC/DC	Acesso a Corpora / Disponibilização de Corpora	pt
ANC	American National Corpus	en
ARCOLEX	Arabic Raw Corpora for Lexical purposes	ar
ArGW	Arabic GigaWord	ar
ArNW	Arabic Newswire	ar
ARTFL	ARTFL-FRANTEXT database	fr
BAC	Buckwalter Arabic Corpus	ar
Birmingham	Birmingham Corpus	en
BNC	British National Corpus	en
BoE1	Bank of English (1993)	en
BoE2	Bank of English (2006)	en

BOLC	Bononia Legal Corpus	it
Brown	Brown corpus	en
BYU	BYU Corpus of American English	en
CEG	Corpws Hanesyddol yr Iaith Gymraeg-A Historical Corpus of the Welsh Language	cy
CIPM	Corpus Informatizado do Português Medieval	pt
CNK	Český národní korpus-Czech National Corpus	cs
ColFIS	Corpus e Lessico di Frequenza dell'Italiano Scritto	it
CORDE	Corpus diacrónico del español	es
CORGA	Corpus de Referencia do Galego Actual	gl
CORIS/CODIS	Corpus di Italiano Scritto	it
CREA	Corpus de Referencia del Español Actual	es
CRPC	Corpus de Referência do Português Contemporâneo	pt
CTG	Corpus Técnico do Galego	gl
CTILC	Corpus textual informatitzat de la llengua catalana	ca
CucWeb	Corpus d'Ús del Català a la Web	ca
Cumbre	Cumbre. Corpus lingüístico del español contemporáneo	es
Davies	Corpus del Español [Davies/NEH/BYU]	es
DeReKo I	Deutsches Referenzkorpus I	de
DeReKo II	Deutsches Referenzkorpus aktuell	de
DWDS-E	Das Digitale Wörterbuch der deutschen Sprache - Ergänzungscorpus	de
DWDS-K	Das Digitale Wörterbuch der deutschen Sprache - Kerncorpus	de
EPG	Ereduzko Prosa gaur	eu
Frantext	Base Textuelle Frantext	fr
ftc	Suomen kielen tekstikokoelma-Finnish Text Collection	fi
Helsinki	Helsinki Corpus of English Texts	en
HNK	Hrvatski nacionalni korpus-Croatian National Corpus	hr
ICE	International Corpus of English	en
ILSP	Hellenic National Corpus - Εθνικός Θησαυρός Ελληνικής Γλώσσας	el
IULACT	IULA-Corpus textual especialitzat plurilingüe	ca
K2000	Korpus 2000	da
KDK	KorpusDK	da
KG	Klasikoen Gordailua	eu
Lexesp	Lexesp. Léxico informatizado del español	es
LIMAS	Limas-Korpus	de
LLELC	Longman/Lancaster English Language Corpus	en
LOB	Lancaster -Oslo/Bergen Corpus	en
MNSZ	Magyar Nemzeti Szövegtár-Hungarian National Corpus	hu
NChÉ	Nua-Chorpas na hÉireann-The New Corpus for Ireland	ga
OC	Oslo Corpus of Tagged Norwegian Texts	no
OEHTC	Orotariko Euskal Hiztegiaren Testu Corpusa	eu
PELCRA	Polish and English Language Corpora for Research and Applications	pl
RNC	Russian National Corpus - Национальный корпус русского языка	ru
SB	Språkbanken-The Swedish Language Bank	sv
SCOTS	SCOTS Project - Scottish Corpus of Texts and Speech	en
SNK	Slovenský národný korpus-Slovak National Corpus	sk
SYN1	SYN2000	cs
SYN21	SYN2005	cs
TMILG	Tesouro Informatizado da Lingua Galega	gl
XXMECE	XX. mendeko euskararen corpus estatistikoa	eu
ZTC	Zientzia eta Teknologiaren corpusa	eu

[1] Sortzaileen ikuspegian, hizkuntza aztertzeko eta teoria linguistikoa eraikitzeko, aski zen “hiztun idealak gramatikaltzat joko lituzkeen perpausak” aztertzea, eta hori introspekzioaren bidez egin lezake ikertzaileak berak (Rojo 2002). Hizkuntzalariek aztertu beharreko datuak ez ziren hizkuntza-erabilera errealean jasotakoak: horiek performantziaren emaitzak dira, eta gaitasun edo *competence* delakoa da linguistikaren aztergaia euren iritziz.

[2] http://www.euskara.euskadi.net/r59-iktcont/eu/contenidos/informacion/ikt_inbentarioa/eu_tic/ikten_inbentarioa.html [2008-03-31].

[3] http://sli.uvigo.es/CLUVI/corpus.html [2008-03-31]

[4] http://www.ehu.es/kontsultak/itzulpen/KP2.HTM [2008-03-31]

[5] XXMECEn, ahozko hizkuntzako laginak daude, ahozkoa transkribatu eta argitaratu denean betiere.

[6] *SpeechDat* proiektua (http://www.speechdat.org/ [2008-03-11]), Europar Batasunak finantzatua. Telezerbitzuak eta ahots bidezko interfazeak garatzeko hainbat hizkuntzatako datu-baseak eratzea du helburu, eta euskara da hizkuntzetariko bat; ELDA bitartez bakarrik jakin daiteke berorren berri, baina hala ere oso positibotzat hartu behar da euskara landu izana.

[7] Ez ditugu arakatu banakoek egindako corpusak (dela tesi-lanetan dela bestelako ikerketetan garatuak), Jon Askerena izan ezik, ELDAko katalogoan salgai aurkitu dugulako eta unibertsitate bati atxikita dagoelako.

[8] Ikerketarako, corpusa doan eskuratu ahal izango da, hitzarmena sinatuta; komertzialki ustiatzeko, lizentziapean.

[9] http://www.webcorp.org.uk/ [2008-03-31]

[10] http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?sprache=en[0] [2008-03-31]

[11] http://www.kwicfinder.com/KWiCFinder.html [2008-03-31]

[12] http://borel.slu.edu/crubadan/stadas.html [2008-03-31]

[13] www.corpeus.org [2008-03-31]

[14] Corpus konparagarriak bi hizkuntza edo gehiagotako testuez osatuak dira, eta, corpus paraleloetan ez bezala, testuak ez dira bata bestearen itzulpenak. Konparagarriak direla esaten da irizpide baten edo batzuen arabera 'antzekoak' direlako (gaiak, datak, diskurtso-mota…). Euskaraz eta beste hainbat hizkuntzatan, zaila da hizkuntza-bikote batzuen corpus paraleloak lortzea, eta corpus konparagarriak baliabide interesgarriak dira, esaterako, baliabide lexikal edo terminologiko elebidunak eratzeko (horretarako, ordea, tresnak behar dira; hori da Elhuyar Fundazioaren *AzerHitz* proiektuaren helburua).

[15] Esaterako, corpus orekatuetan, beharrezkoa izaten da unibertsoaren inbentario batetik abiatuta lagintze-eredu bat erabiliz obren zozketa egitea, eta corpuserako hautatutako obretatik ere lagin bat hartzea. Corpusa orekatua ez bada (oportunista baizik), urrats hori ez da beharrekoa. Bestetik, corpus batzuetan formatu estandar bat erabiltzen da. Estandar ezagunenak TEI (*Text Encoding Initiative*) (Ide & Véronis 1995) eta CES (*Corpus Encoding Standard*) (Ide 1998). Beste batzuek berariazko SGML edo XML instantziazioak edo testu hutsa ere erabiltzen dituzte. Horrez gain, batzuetan egitura-etiketatzea areago aberasten da, eskuz edo erdi automatikoki.

[16] Bertsio irekia du, *IMS Open Corpus Workbench*: http://cwb.sourceforge.net/ [2008-03-31]

[17] http://www.oucs.ox.ac.uk/rts/xaira/ [2008-03-31]

[18] http://www.ddc-concordance.org/ [2008-03-31]

[19] http://philologic.uchicago.edu/ [2008-03-31]

[20] http://www.sketchengine.co.uk/ [2008-03-31]

[21] Honela definitu du J. Sinclairrek: "A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials." (Sinclair 1996). Euskaraz, honela definitu du A. Sagarnak: “Hizkuntza bati buruzko ahalik eta informaziorik osatuena emateko prestatuta dagoen corpusari erreferentzia-corpusa esaten zaio. Hizkuntzaren ahalik eta aldaera gehienen berri emateko, behar den adinako tamaina izan behar du. Tamaina garrantzitsua da, baina, gainera, eredu egoki baten arabera hautatutako testuak eduki behar ditu erabileremu, genero, erregistro eta gainerako bereizgarriak kontuan hartu eta orekatua izateko. Hartara, erabil daiteke gramatikak, hiztegiak, thesaurusak eta beste hainbat tresna lantzeko.” (Sagarna 2007).