Itzulpen-corpusak hiztegigintzan: baliabide eta bide berriak

David Lindemann
Itzultzailea eta Euskal Herriko Unibertsitatean doktoregaia

1. Sarrera

Egun, inork ez du zalantzan jartzen corpus hizkuntzalaritzak nabarmen aberastu duela hiztegigintza (ikus gaiaren oinarrizko aurkezpenak Atkins & Rundell 2008; Svensén 2009). *Corpus revolution* deitutako berrikuntzari esker (Rundell & Stock 1992; Krishnamurthy 2002), hizkuntzaren erabilera hobeto deskriba daiteke gaur. Corpus garaiaren aurreko hiztegiek, zenbait egilek nabarmendu dutenez (Klosa 2007), hitzen adiera (eta hiztegi elebidunez ari garenean, adieraren itzulpen-ordain) bakan edo arraroak aipatzen dituzte, eta baztertu egiten dituzte, aldiz, garrantzi edo maiztasun handikoak. Orain, maiztasun neurketak hiztegigintzaren tresneriaren ohiko osagai bihurtu dira, eta corpus datuetan oinarritutako hitzen profilek (*word sketches, Wortprofile*) hitzaren agerkideei eta erlazio sintaktikoei buruzko ebidentzia ematen digute (Kilgarriff & Tugwell 2002). Maiztasun datuak, hitzaren agertokietako testuinguruak islatzen dituzten konkordantziak eta hitz-profilak segundo batean baino denbora gutxiagoan sor eta ikusaraz daitezke, dagokion hizkuntzan testu-corpus behar bezain handiak eta hizkuntzaren prozesamendurako tresnak eskuragarri izanez gero.

Hiztegian erabilera-adibideak emateko, aurreko garaietan eskuz batu behar izaten ziren aipuak, zegokien lemaren arabera *zapata kaxa* famatuetan gorde eta hiztegia editatzean ateratzen zirenak. Corpusei esker, lipar batean sor ditzakegu hitz baten inguruko testuinguru edo konkordantziak, hiztegigintza-lanetan erabiltzeko. Baina hiztegigileari hiztegi baten edukia editatzen laguntzeaz gain, beste zeregin bat betetzen dute corpus-datuek hiztegigintza arloan: gero eta sareko

hiztegi gehiagotan, erabiltzaileari erakusten zaion hiztegi sarrerari (*editoriala* eta *estatikoa* dei diezaiokegun horri) corpus datuetan oinarritutako edukiak gehitzen zaizkio, erabiltzaileak bilaketa gauzatzen duen momentuan sortuta, hala nola corpusetatik erauzitako adibide-esaldiak, bilatutako hitza dutenak. Hiztegi elebidun batentzat prestatutako datu-sortak direnean, esaldi bakoitzak dagokion itzulpena du alboan; hau da, konkordantzia edo KWIC *(keyword in context)* elebidun paraleloa. Paragrafo- edo esaldi-bikote elebidunak itzulpen-corpus batetik erauzi ahal izateko, corpusa lerrokatu behar da, hots, zein segmentu zein segmenturen itzulpena den zehaztu. Paragrafo edo esaldi mailan lerrokatutako itzulpen-corpusari, corpus paraleloa (*parallel corpus*) deitzen diogu.

Baliabide askoko hizkuntza *handietan*, alemanez, esaterako, hiztegigileak, baita beste edozein erabiltzailek ere, corpusetatik erauzitako datu-sortak ematen dituzten maila goreneko baliabideak ditu eskuragarri, DWDS hiztegi-atariaren eskaintza, kasu.[1] *Corpus hiztegigintza* batek zer eragin duen hiztegigintza lanetan eta corpusetako datuek zenbateraino aberastu ditzaketen sareko hiztegi baten edukiak, horra hor artikulu honetan interesatuko zaizkigun galderak.

2. Corpus elebakarrak hiztegigintza elebiduneko lanetan

Gure interesgune nagusia den hiztegigintza elebidunean murgildu baino lehen, ikus dezagun labur zer ekarpen zor diogun corpus hiztegigintzari ele bakarrari dagokionez. Alde batetik, hiztegi baten lehmategia eraikitzeko lanak aipatuko ditugu, eta bestetik, lema bakoitzari dagozkion entitate sintaktikoak (*kategoria gramatikalak*) definitzekoak. Bi kasuetan, corpusetatik erauzitako maiztasun-datuak dira lanean erabiltzeko ebidentzia ematen digutenak. Horretaz gain, jakina, erabilera-adibideak dira corpusek ematen dituzten datu garrantzitsuak, euskara ele bakarrarentzat corpusetan oinarritutako *Egungo Euskararen Hiztegia*k[2], esaterako, baliatzen dituenak. Honetara hurrengo atalean itzuliko gara, hiztegigintza elebidunari begira.

Hitzen erabilera-maiztasuna eta hitz haiei dagozkien hiztegi-sarreren ikustaldi-kopurua erlazionatuta daude neurri handi batean (De Schryver et al. 2010; Wolfer et al. 2014). Horrek esan nahi du jokabide zentzuduntzat jo dezakegula hiztegiaren lehmategia eraikitzeko lanak corpusetan oinarritutako maiztasun-zerrenda batetik abiatzea, hiztegi berri bat sortzean, bederen: erabiltzaileak maiz agertzen diren hitz hauek bilatuko ditu hiztegian. Aldi berean, datuak eskura edukiz gero, lemari lotutako maiztasun-informazioa eman ahal zaio erabiltzaileari hiztegi-sarreran bertan, ingelesezko hizkuntza-ikasleentzako hiztegigintzan aspalditik egiten den bezala (Kilgarriff 1997). Alemanez, esaterako, badaude era profesional batean hiztegigintzari lehengaia eskaintzeko asmoz sortutako maiztasun-zerrendak.[3] Euskaraz, maiztasun hiztegiak

1. http://dwds.de/. Atari honek, puntako baliabide eta teknologietan oinarrituta, alemanezko zenbait hiztegiren edukia, corpus konkordantziak eta hitz-profilak eskaintzen ditu, erabiltzaileak bere interesaren arabera molda dezakeen modu batean.
2. Ikus http://www.ehu.eus/eeh/.
3. Ikus http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html. Maiztasun-zerrenda guztiak lizentzia librekoak dira.

argitaratu dira, corpus batean maizen agertzen diren hitzen bildumak.[4] Guk ere proposatua dugu lehenago ere zer metodo erabil daitekeen eskuragarri ditugun euskarazko corpus handienetatik erauzitako maiztasun-zerrendetan oinarritutako lemategi bat sortzeko (Lindemann & San Vicente 2015). Euskarazko hiztegietan sarrera izanik ere egungo corpusetan agertzen ez diren lema-multzoak zehaztu ditugu, baita alderantziz ere: egungo corpusetan bai, baina gaur arteko hiztegietan agertzen ez zaizkigunak. Corpusetan agertzeak hitzaren erabilera egiaztatzen duela onartzen badugu, hutsetik abiatzen den hiztegigintza-asmo batentzat garrantzi handiko ebidentziatzat jotzen dugu bietan agertzea: hiztegietan eta corpusetan. Batez ere euskarazko web-corpusek lexema eta adiera berri asko dituztela azpimarratzen du Leturiak (2014). Corpusetan bai, baina hiztegietan (oraindik) agertzen ez diren euskarazko hitz haiek, lema bihurtzeko hautagai direnak, web-corpus batek dituen datu-masa erraldoian ezkutatzen dira, eta ez da erraza haiek iragaztea: eskuzko lana ezinbestekoa da zeregin honetan.

Bestalde, lemak entitate sintaktiko gisa identifikatzeko ebidentzia eskaintzen digute corpusek. Izan ere, corpus bateko edukiak prozesatzeko pauso bat hitz guztien etiketatze morfosintaktikoa da, *Lemmatizer-PoS-Tagger* deituriko tresna batez gauzatzen dena. Prozesatze horren funtzioa bikoitza da: alde batetik, testu-hitz bakoitzari dagokion forma kanonikoa zehaztea, hots, hiztegian sarrera-buru izan daitekeen lema zehaztea; eta bestetik, hitzaren kategoria (eta azpikategoria) gramatikala definitzea. Ikus dezagun adibide bat, euskarazko *alegia* formari web-corpus handi batean[5] zehar *EusTagger* etiketatzaileak (Aduriz et al. 1996, 2007) esleitu dizkion kategoriak erakusten dituena:

1. taula: Elh200 corpusetik erauzitako maiztasun datuak *alegia* lemarentzat azpikategoria mailan

posizioa maiztasun-zerrendan	**agerpen kopurua**	**kategoria gramatikala**
618	41.106	lokailua
3.882	4.208	izen arrunta
10.407	921	izen berezia: leku izena
1.440.023	1	adjektiboa
1.880.968	1	adberbioa

Agerpen kopururako atalase minimoa finkatzen badugu,[6] *alegia* lemari dagokion hiru hiztegi-sarrera (edo hiru ataleko sarrera bat) izan beharko ditu hiztegiak, datu horien arabera:

4. Hiru argitalpen ditugu: 1977ko corpus batean oinarritutako Sarasolaren (1982) lana, ahozko corpus batetik abiatzen den Etxebarria eta Mujikarena (1987), eta, hirugarrenik, UZEI (2004) maiztasun hiztegia, *XX. Mendeko Corpusa* oinarria duena.
5. Datu hauek Elhuyar200 web-corpusetik erauzi ditugu. Corpus horren garapena Igor Leturiari (2014) zor zaio.
6. Sinclair (2005) lanean 20 agerpen proposatzen dira, hiztegigintza lanetan esanahia definitzeko ebidentzia nahikoa izateko. Etiketatzaile morfosintaktikoak *alegia* forma kasu banatan adjektibo eta adberbio gisa etiketatu izana, beraz, ez da nahikoa *alegia* lemarentzat kategoria hauetako hiztegi-sarrerak egiteko, etiketatze hori zuzena ala okerra (prozesamendu errorea) den gorabehera.

alegia lokailuarena, *alegia* izen arruntarena, eta *alegia* (*Alegia*) herri izenarena. Ikusten dugunez, *EusTagger* tresnak maiztasun-datuak ematen dizkigu azpikategoria gramatikalari dagokionez (izen arrunta, izen berezia, etab.). Entitate sintaktiko bakoitzerako, beraz, maiztasun-datuak ditugu, hiztegi-sarrera batean ere eman daitezkeenak.

3. Corpus paraleloak eta hiztegigintza

3.1. Euskararekiko corpus paraleloak: artearen egoera

Elhuyar Fundazioaren web-corpusen atalean euskara-gaztelaniazko corpus paraleloa dago kontsultagai.[7] Corpusa automatikoki osatu da, eduki elebiduna duten domeinuak sarean bilatuz, eta domeinu horietatik elkarren itzulpena diren esaldiak erauziz. 18.753.613 testu-hitz ditu (7.891.104 euskaraz eta 10.862.509 gaztelaniaz). 659.630 segmentu elebidun ditu. Egungo euskara-gaztelaniako corpus paralelo handiena da, eta edonork erabil dezake bilaketak egiteko interfazea.

Euskal Herriko Unibertsitateko Euskara Zerbitzuak 2002. urte ingurutik argitaratzen dituen ikasliburuen itzulpenez osatua da *EHUskaratuak* corpus paraleloa. Giza zientzia, gizarte-zientzia, ingeniaritza eta teknologia, osasun-zientzia eta zientzia zehatz eta materiaren zientzia arloetako itzulpenak biltzen ditu, ingelesetik, frantsesetik eta gaztelaniatik burutuak; corpus paralelo eleaniztuna da, beraz. Sareko interfaze bat erabilgarri dago UPV/EHUko webgunean.[8] Corpusa Elhuyar Hizkuntza eta Teknologiak garatu du, eta itzulpen berriak gehituko zaizkio corpusari etorkizunean.

3.2. Euskara-alemanezko corpus paraleloak

Euskara-alemana hizkuntza-bikoteari bigarren mailakoa dei diezaiokegu, euskararen barne-erdarak (gaztelera eta frantsesa) edo kanpo-erdara nagusia (ingelesa) kide diren bikoteen atzean, hiztegigintza nahiz corpusgintzari begira. Atal honetan ikusiko dugunez, bigarren mailako bikote honetan kalitatezko corpus paraleloak sortzeko ahaleginak egin berri dira. Bigarren maila honetako beste bikote guztietan, berriz, egiteke geratzen zaizkigu oraindik, guk dakigunez.

EHUko Letren Fakultatean kokatutako TRALIMA/ITZULIK ikerketa taldean, alemanezko literatura-lanak nahiz euskarazko itzulpenak biltzen dituen corpus paralelo bat sortu du Uribarri, Sanz eta Zubillaga hirukoteak (Sanz et al. 2015). Corpusak itzulpen zuzenak eta zeharkakoak biltzen ditu. Alemanetik zuzenean euskaratu diren testuak alemanezko jatorrizkoan nahiz euskarazko bertsioan daude jasota; zeharka burutu diren itzulpenak, aldiz, jatorrizkoan, gaztelerazko *zubi-bertsioan* eta euskarazkoan. Azken kasu hauetan, beraz, corpusa hirueleduna

7. Ikus http://webcorpusak.elhuyar.org/cgi-bin/kontsulta2.py.
8. Ikus http://ehuskaratuak.ehu.eus/bilaketa/.

da. Corpusa osatzen duten testuetatik, alemanezko 81 testu literario eta euskarazko itzulpen zuzenak izan ditugu eskuragarri, paragrafo edo esaldi mailan lerrokaturik, hiztegigintza arloko gure esperimentuetan erabiltzeko. Euskarazko testuen lematizatzea *EusTagger* tresnaren bidez gauzatu dugu[9]. *SketchEngine* (Kilgarriff et al. 2004) tresnaren bitartez, alemanezko testuen lematizatzea (*TreeTagger*, (Schmid 1995)) burutu eta, testu paralelo guztiak aplikazioan kargaturik, bilaketak egin eta konkordantzia paraleloak sortu ditugu. Horrela sortutako corpus paraleloak bina milioi testu-hitz inguru ditu euskaraz eta alemanez. Bestetik, Bibliaren alemanezko eta euskarazko bertsio bana erabili dugu corpus paraleloa sortzeko.[10] Bibliatik corpus paraleloak sortzea nahiko ariketa erraza da, bertset-kodeen bitartez lerroka baitaiteke. Bibliaren itzulpena hainbat eta hainbat hizkuntzatan, eta, horretaz gain, hizkuntza bakoitzean maiz bertsio batean baino gehiagotan eskuragarri izatea oinarri egokia da era askotako konparaketak egiteko (Resnik et al. 1999; Nida & Taber 2003). Corpusak 700.000 bat testu-hitz ditu hizkuntza bakoitzean.

Eskuz hautatutako iturri literario hauetaz gain, euskara-alemanezko corpus paraleloak modu eraginkor batean osatzeko balioko luketen beste zenbait datu-iturri ere aipa daitezke: ikus-entzunezkoen azpitituluak eta software lokalizazioak.[11]

3.3. Corpus paraleloak hiztegigintza lanetan

EuDeLex izeneko euskara-alemanezko gure egitasmo berriaren garapen prozesuan, aurreko atalean deskribatutako corpus elebiduna erabiltzen dugu. Batetik, konkordantzia paraleloak baliatzen ditugu alemanezko lemaren euskarazko ordainak definitzeko lanetan. Gure esperientziaren arabera, batez ere kontzeptu abstraktuen kasuetan lagungarriak izan dira itzultzaileen proposamenak euskarazko ordain egokienak hiztegi-sarrera elebidunerako hautatzeko, baita dagokion lema duten esapide eta lokuzioak itzultzeko ere. Bestetik, euskara-alemanezko glosarioak zirriborratu ditugu datu haietan oinarriturik: esaldi- edo paragrafo-mailako lerrokaketatik hitz-mailako lerrokaketara heltzen diren tresna automatikoen bitartez, aleman-euskarazko itzulpen-ordainen bikote izateko hautagaiak definitu ditugu (ikus Lindemann et al. 2014).

3.4. Corpus paraleloetatik erauzitako adibide-perpausak hiztegi-sarreraren osagai

Corpus paraleloetatik erauzitako konkordantziak hiztegi elebidunetako sarreretan islatzeko aukera 1990eko hamarkadaren hasieratik aipatu izan den arren (lehenengo aipamenak Atkins 1996; Dickens & Salkie 1996), aukera berri hori gauzatu izanaren zenbait adibide baino ez

9. EHUko IXA taldeko Gorka Labakari zor dizkiogu euskarazko lematizatze eta etiketatze lanak.

10. Elhuyar Fundazioko Xabier Saralegiri zor zaio lan hori, ikus Lindemann et al. (2014).

11. OPUS egitasmoak euskarazko eta beste hainbat hizkuntzatako datu paraleloak biltzen ditu, iturri haietatik batik bat (Tiedemann 2012). Euskara-alemanezko konbinazioan, adibidez, OPUS datu-basean batutako testu paraleloa aintzat hartzeko tamaina batera irits liteke datozen urteotan.

ditugu egun arte. Gorago esan bezala, corpusak hitzen adiera nagusiak edukitzen ditu, hiztegi sarrera editorial batean maiztasun gutxiagoko adieren artean *ezkutaturik* gera litezkeenak. Bestetik, hitz baten adiera helburu hizkuntzan islatzeko, itzultzaileek maiztasun gutxiko ordainak hautatzen dituzte maiz, hiztegi-sarrera editorial batean agertzen ez direnak. Wolfgang Teubert autorearen hitzetan, "the translator's design space is much larger than the language-neutral conceptual ontology (or the traditional bilingual dictionary) would leave us to believe" (Teubert 2002). Gainera, itzultzaile bat ez da mugatzen hitz bat hitz batez ordezkatzera; aitzitik, egitura sintaktikoa ere moldatu, laburbildu, garatu, edo helburu hizkuntzan egokia gertatzen den esapide batez itzul dezake, esaldi osoa modu egoki batean islatzeko bere arduraren arabera. Giza itzultzaileak baino sor ez ditzakeen moldapen horiek dira kalitatezko corpus paralelo batek eskaini dituen altxorrak.

2. taula. Corpus paralelotik erauzitako alemanezko *Ärger* lemaren konkordantziak

Konkordantzia DE	*Konkordantzia EU*	*Transformazio sintaktikoa*
Aber auf dieser Route **hatte** ich noch nie **Ärger** mit ihnen!	Baina bide honetan ez **dut** inoiz haiekin **arazorik izan**!	V+N > V+N
Nun war der **Ärger** in Capricorns Stimme nicht mehr zu überhören	Dagoeneko nabaria zen Kaprikornioren ahotsean **haserrea**.	N > N
Meggie konnte nicht verhindern, dass ihre Stimme **vor Ärger** zitterte.	Meggiek ezin ekidin zezakeen ahotsak **amorruz** dar-dar egin ziezaion.	prep.+N > N+postp.
Meggie sah, wie sich Bastas Schultern **vor Ärger** spannten.	**Haserrearen haserrez**, Bastaren sorbaldak tenkatu egiten zirela ikusi zuen Meggiek.	prep.+N > (esapidea) adb.
Mensch, biste nicht froh, daß du den ganzen **Ärger** los bist mit den Weibern?	Eta hi zer, ez al hago pozik, andreekin hituen **saltsa** guztiak bukatuta?	N > N
Bastas Lippen wurden schmal **vor Ärger,** doch er verkniff sich eine Antwort und…	Bastak ezpainak estutu zituen **haserre**, baina erantzuteko gogoari eutsi zion.	prep.+N > adb.

Egokitzapen ezberdinak ikusten ditugu goiko taulako euskara-alemanezko adibideetan. Corpus paraleloetako datuetan, beraz, itzultzaileek jorratutako bideak azaltzen zaizkigu, itzultzaileak zer prozeduraz baliatu diren, kasuz kasu, testuingurua aintzat harturik; eta horixe da hiztegi-sarrera elebidun batek bere estutasunean orain arte nekez aurreikus eta isla zezakeena. Horrelako datuetan hizkuntzalari konputazionalek bilatzen dute itzulpen-makinek oraindik imitatzen ez dakitena, taulan islatu duguna, "syntactically motivated transformation rules that

explain human translation data" (Galley et al. 2004). Testuingurua duten itzulpenez, corpus paraleloz elikatzen dituzte estatistika hutsa baliatzen besterik ez dakiten itzulpen-tresna automatikoak, haiei giza itzultzaileen jokabideak *ulertarazteko* ahaleginean.

Hiztegi-sarrera editorial batean, lemak entitate sintaktiko gisa dituen izaerak (hau da, haren kategoria gramatikalak) mantendu ohi dira hizkuntza batetik bestera, helburu hizkuntzan posible den heinean, bederen.[12] Beraz, hiztegi-sarrera elebidunarekin batera benetako adibideen ondoan benetako itzulpenak ematea lagungarria da benetan erabiltzen den hizkuntza islatzeko, eta corpusetako konkordantzia paraleloak horretarako baliatzea jokabide egokia izan daiteke. Zalantzarik gabe, corpus paraleloaren edukien kalitatea aldagai garrantzitsua da; orain arteko gure esperientziaren arabera, liburuen itzulpenak dira helburu horretarako emankorrenak.

3.5. Konkordantzia paraleloak hiztegi elektronikoetan: artearen egoera

Itzulpen-corpusetatik erauzitako datuak hiztegi-erabiltzaileari zuzenean erakustea eta horrela erabiltzailearen galderari benetako erabilera-adibide eta haien itzulpenez erantzutea joera berria dugu hiztegigintza elebidun elektronikoan. Bide berriak jorratzen dituzten bi hiztegiatari aipatuko ditugu.

Linguee izeneko egitasmoa 2007an sortu zen, eta 2009tik aurrera, erabiltzaile-interfazea erabilgarri dago edonorentzat.[13] 2015ean, 25 hizkuntzatako datuak biltzen ditu eta haien arteko oinarrizko hiztegi-sarrera elebidunak eta itzulpen-adibideak eskaintzen ditu. Itzulitako esaldien datu-baseak Interneten aurkitutako testu paraleloak ditu. Ikus dezagun adibide bat, gaztelerazko *falda* lema polisemikoaren ingelesezko itzulpenak biltzen dituena (1. irudia):

Ikusten denez, galderaren erantzunak biltzen dituen orri honetako goiko partean, hiztegi-sarrera elebiduna dator, gaztelerazko hitzaren ezaugarri morfosintaktikoak eta hiru adierentzako ingelesezko ordain bana ematen duena. Horretaz gain, hitzak entzuteko aukera ematen duten botoiak, eta *falda* parte duten unitate lexiko hitz anitzekoen ingelesezko ordainak ematen dira *adibide* gisa. Azpian, ingelesezko nahiz gaztelerazko *wikipedi*ek ematen dituzten definizio laburrak ikus daitezke, dagozkien gezietan klikatuz. Honaino, beraz, oso eskaintza oinarrizkoa da ikusten duguna.

Hortik behera, itzulitako adibide-esaldiak datoz, eta irudi honetan haien parte txiki bat baino ez dugu islatu. Ikusten denez, gaztelerazko hitza nahiz ingelesezko ordaina nabarmentzen dira, bilatutako hitza eta ordainak testuinguruetan zehar bizkorrago ikusteko. Honelako esaldi-bildumen balioa bistakoa da: hiztegi-erabiltzaileak bere kasuari hurbila zaion testuingurua bi-

12. Batzuetan, helburu-hizkuntzan posible dena azken muturreraino erabiltzen da, kategoria gramatikala mantentzearren. Horrela, *Euskal WordNet* datu-base lexikalean ingelesezko hainbat izenetako ordainak eratorritako formak dira: adibidez, hainbat *-t(z)ea* eta *-tasun* izen agertzen dira euskarazko ordain gisa, ingelesezko iturria izena denean. Horietako asko ez dira eratorritako forma horretan agertzen euskarazko hiztegi-lemategietan: *connivance, secret approval, tacit consent*: 'ados jartze', *spraying, crop-dusting*: 'aerosolez bustitze', *randomness, stochasticity*: 'aleatoriotasun'.

13. Ikus http://www.linguee.com.

falda con vuelo *f* — fit and flare skirt *s*

▶ Wikipedia

▼ Fuentes externas

Godet: pliegues de tela cortados a capa que se añaden en la parte baja de la falda para darla más vuelo y más peso. esflamenco.com	Godet: folds of cloth cut i to give it more fullness a
La segunda figura lleva falda y sostiene algo que podría ser una bolsa de red. paracas.se	The other figure is weari
Incluso antes de que los bebés puedan sentarse por sí mismos, pueden acurrucarse en su falda y mirar imágenes de libros con dibujos en blanco y negro o fotos de rostros. es.4children.org	Even before babies can and look at board books
Corte y deshuesado: separar toda la falda del cuarto trasero desde el corte recto a la altura de la octava costilla, cortando [...] eur-lex.europa.eu	Cutting and boning: rem hindquarter by a cut from
Sí, tenemos un pantalón corto o una falda en beis, una camiseta roja y una camisa azul. fimcap.org	Yes, a beige trouser or s
La mujer lleva una blusa blanca y una falda negra. resources.rosettastone.com	The woman is wearing a
La villa está enclavada junto al río Matarraña, en las proximidades del límite de provincia de Tarragona, en la falda de la Reserva Natural de los Puertos de Beceite. expozaragoza2008.es	The town is right next to province of Tarragona, o Beceite.
Asimismo, las autoridades han prohibido a las mujeres usar bicicletas (un vehículo esencial para poder comerciar) y las han obligado a llevar falda. daccess-ods.un.org	The authorities have als key vehicle for access to
El pistón de falda corta se mueve en un cilindro recubierto de un compuesto cerámico que reduce las pérdidas friccionales, [...] yamaha-motor-europe.com	The short skirt piston run reduces frictional losses,
Alojamiento en suites con piscina o pabellones de caprichosa arquitectura en la falda de cuatro volcanes humeantes. guiamejordelmundo.com	Accommodation in suite capricious design set on
En el armario, todavía están la blusa y la falda del primer día, nada más. basqueliterature.com	The jacket and skirt from wardrobe.

1. irudia: *Linguee* hiztegiaren lagina

latuko du, edota kasuz kasu nola itzuli den ikusi. Jakin nahi duena, beharbada, oso galdera berezitua da, termino gisako erabilera bat, ibilgailu-motorreko *pistón de falda corta* ingelesez *short skirt piston* ote den, adibidez, eta datu-base terminologiko egokia eduki ezean, hemen bai aurkitzen du erantzun bat, kasualitatea dei badiezaiokegu ere. Gaztelera-ingelesa konbinazioan badira banku terminologikoak kasu honetarako[14]; baliabide gutxiagoko hizkuntza-bikote baterako, aldiz, corpus paraleloak baliatzen dituen tresna bat edukitzea bereziki lagungarria dela iruditzen zaigu.

1.: ya que	2.: sin embargo	3.: por lo tanto	4.: además	5.: por otro lado
6.: aunque	7.: es decir	8.: a partir de	9.: llevar a cabo	10.: por lo que
11.: asimismo	12.: no obstante	13.: por otra parte	14.: a continuación	15.: aprovechar
16.: actualmente	17.: a pesar de	18.: so	19.: así como	20.: mediante
21.: en cuanto a	22.: realizar	23.: aportar	24.: sobre todo	25.: cumplir
26.: a su vez	27.: debido a	28.: en este sentido	29.: por su parte	30.: atentamente
31.: presupuesto	32.: conseguir	33.: seguimiento	34.: destacar	35.: siempre y cuando
36.: tener en cuenta	37.: de esta manera	38.: vigente	39.: por tanto	40.: posteriormente
41.: con respecto a	42.: facilitar	43.: funcionamiento	44.: empresa	45.: convocatoria
46.: sino	47.: si bien	48.: propuesta	49.: compromiso	50.: resumen

Linguee gunean bertan ematen diren ikustaldi handiko sarreren zerrenden arabera (ikus goiko zerrenda), ondorengoa bezalakoak dira oso maiz baliabide honen bitartez argitzen saiatzen diren kasuak: hizkuntza batek berezko dituen hitz anitzeko esapideak, gaztelerazko *en cuanto a* bezalakoak. Ikus ditzagun haren itzulpenak:

En cuanto a la composición demográfica del país, el proceso de envejecimiento de la población impone un reto a la política social. daccess-ods.un.org	With regard to the country's demographic composition, the ageing of the population poses a challenge to social policy. daccess-ods.un.org
En esta labor se debe tener en cuenta que persisten las diferencias entre los países en cuanto a su capacidad para la difusión de información. daccess-ods.un.org	Those initiatives must take into account the differences that persisted among countries' capacities to disseminate information. daccess-ods.un.org
En cuanto a la cuestión de la energía, por desgracia soy bastante pesimista. europarl.europa.eu	When it comes to the energy issue, I am unfortunately rather pessimistic. europarl.europa.eu
Dado que la búsqueda se inicia en cuanto empieza a escribir, si por ejemplo escribe "B", todos los archivos con nombres [...] windows.microsoft.com	The search begins as soon as you begin typing-so if you type "B," for example, all the files with names starting with the [...] windows.microsoft.com
En la actualidad, ellos deciden en cuanto a las ofertas sobre la base de la calidad y el precio. gopacnetwork.org	At present, they decide on bid offers using the bases of quality and price. gopacnetwork.org
[...] Unidos se encuentra mucho más adelantado que el resto de las naciones en cuanto al gasto en servicios de salud per cápita, ocupa el lugar 48 en expectativa de vida, dijo. forumfed.org	At the same time, the United States remains far in front of all other nations in per capita health care spending while it ranks 48th in life expectancy, she said. forumfed.org
En cuanto a mí, no puedo pensar en hacer películas en las que no me involucre, prefiero esperar un proyecto que me estimule de verdad. cineuropa.mobi	It is impossible for me to get involved in films that I don't like so I just wait for a project that really tickles my fancy. cineuropa.mobi
[...] diversifica la oferta de educación se observan desigualdades de una magnitud sin precedentes en cuanto al acceso y la calidad. unesdoc.unesco.org	But as educational demand increases and supply diversifies, disparities of unprecedented proportions can be observed in respect of access and quality. unesdoc.unesco.org
El parlamento es útil a la burguesía en cuanto logra crear la ilusión de que los obreros eligen a quienes los malgobiernan. ibrp.org	Parliament is useful to the bourgeoisie in that it gives the illusion that workers choose who is to misrule them. ibrp.org

14. Egun, Europar Batasuneko hizkuntza ofizialak batzen dituen IATE datu-base terminologikoa puntako baliabidea dugu; ikus http://iate.europa.eu.

Ikusten eta espero dugunez, *en cuanto a* esapideak askotariko ordainak ditu ingelesez, testuinguruaren arabera. Ordain bakar bat ez da errepikatzen. Kasu batzuetan, ohiko hiztegi-sarrera batean aipatuko litzatekeen ordain zuzenik ez da, bigarren adibidean duguna, kasu: genitibozko egitura baten bitartez itzultzen da gaztelerazko esapidea. Aipatzekoa da, halaber, gaztelerazko laugarren adibideko *en cuanto* egiturak, grafiaren hurbiltasuna gorabehera, ez duela zerikusi semantikorik bilatutako *en cuanto a* esapidearekin; horra hor atari honetan erabilitako metodoaren muga bat. Erabiltzailearentzako onura, berriz, agerian dago: ohiko hiztegi-sarrera batean agertuko ez litzaizkiokeen itzulpenak, testuinguru ezberdinak dituzten itzulpenak datozkio hemen.

Bestetik, maizen gauzatzen diren bilaketa hauek ikusita, ondorengo susmoa sortzen zaigu: hitz anitzeko unitate ugari dira zerrenda hauetan; erabiltzaileek hitz anitzeko unitate haiek beste hiztegi batzuetan, testuingurua duten itzulpenak ematen ez dituzten hiztegietan, aurkitzea espero ez dutelako ote? Hiztegien erabileraren inguruko ikerketa-galdera irekia dugu hemen. *Linguee*, behintzat "a translators' favourite" bilakatu da epe motz batean (Kilgarriff 2013:92).

Linguee atariaren garatzaileek haien metodoen berri ematen ez badute ere, bistan da logaritmoak ere erabiltzen direla corpusak dituen adibide-esaldi guztien artean kontsulta bati erantzuteko egokienak hautatzeko, ez baitago esaldi motzegirik edo luzeegirik ezta oso testuinguru edo hitz bitxia daukan esaldirik. Adibide *onak* hautatzeko logaritmo automatikoak garatu izan dira, eta zenbait kasutan publikoak dira haiek erabiltzeko arautegiak, Kilgarriff (2008) adibide.

Glosbe hiztegi-ataria[15] euskarazko esaldiak eta hiru erdara nagusiez gainerako hainbat hizkuntzatako itzulpenak eskaintzen dituen baliabide bakarra da egun, guk dakigunez. Izan ere, hizkuntza-bikote askorentzako itzulpenak ditu atari honetako datu-baseak. Euskaratik edo euskarara ehun bat hizkuntzatara ditu loturak, eta arlo jakin batzuetarako baliabide baliotsutzat jo dezakegu dagoeneko, informatikarekin zerikusia dutenetarako, bereziki. Izan ere, software lokalizazioetako edukietan oinarritzen dira *Glosbe* datu-basearen euskarazko laginak batez ere.

4. Ondorioak

Itzulpen-corpusak hiztegigintzan baliatzea aberasgarria da, zalantzarik gabe, hiztegigilearentzat nahiz hiztegi erabiltzailearentzat: corpusak tamaina eta osaeraren arabera eskaini ditzakeen adibide eta itzulpenek ohiko hiztegi elebidunen egitura aberasten dute. Adibide-esaldiak eskuz hautatzea eta dagokien hiztegi-sarreran txertatzea, corpusen garaiaren aurretik egiten zen legez, lan erraldoia izaten zen. Gaur, corpus-konkordantziek errazten diote hiztegigileari eduki egokienak hautatzeko lana. Ibon Sarasolaren *Egungo Euskararen Hiztegia* corpus elebakar batetik eskuz hautatutako erabilera-adibideak baliatzearen etsenplu bikaina da. Adibide egokienak eskuz aukeratzeko baliabiderik ez dagoenean edo ahalik eta adibide gehien eman nahi direnean, erabiltzaileak bilaketa egiten duen momentuan sorturiko corpus-konkordantziak erants

15. Ikus http://eu.glosbe.com/eu/.

dakizkioke hiztegi-sarrera estatiko bati, goian aipatutako DWDS atarian gertatzen den legez. Corpus paraleloetatik automatikoki erauzitako itzulpen-adibideak hiztegi elektroniko baten sarreretan zuzenean txertatzea joera berria da; eskuz landu gabeko hizkuntza-bikote askotan, corpus paraleloetan oinarriturik automatikoki sortutako hiztegiak dira eskuragarri dauden baliabide bakarrak. Dena den, kalitatezko corpus hiztegigintza batean ezin dugu ahaztu eskuzko lanaren premia, metodo automatikoen bitartez sortutako emaitza desegokiak iragazteko eta hutsuneak atzeman eta betetzeko orduan, batik bat.

5. Aipamenak

ADURIZ, I., ALDEZABAL, I., ALEGRIA, I., ARTOLA, X., EZEIZA, N. & URIZAR, R. (1996). EUSLEM: A lemmatiser/tagger for Basque. In *Proceedings of EURALEX 1996*. Göteborg: Göteborg University, pp. 17–26.

ADURIZ, I., ALEGRIA, I., ARRIOLA, J., ARTOLA, X., DÍAZ DE ILARRAZA, A., EZEIZA, N., GOJENOLA, K. & MARITXALAR, M. (2007). Different Issues In The Design Of A Lemmatizer/Tagger For Basque.

ATKINS, B.T.S. (1996). Bilingual dictionaries: Past, present and future. In *Proceedings of EURALEX 1996*. Göteborg: Göteborg University, pp. 515–546.

ATKINS, B.T.S. & RUNDELL, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

DICKENS, A. & SALKIE, R. (1996). Comparing Bilingual Dictionaries with a Parallel Corpus. In *Euralex'96 Proceedings*. Göteborg: Göteborg University, pp. 551–559.

ETXEBARRIA, J.M. & MUJIKA, J.A. (1987). *Euskararen oinarrizko hiztegia : maiztasun eta prestasun azterketa*. Gasteiz: Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia.

GALLEY, M., HOPKINS, M., KNIGHT, K. & MARCU, D. (2004). What's in a translation rule. In *Proceedings of HLT/NAACL*. Boston, pp. 273–280.

KILGARRIFF, A. (1997). Putting frequencies in the dictionary. In *International Journal of Lexicography*, 10, 135–155.

KILGARRIFF, A., HUSAK, M., MCADAM, K., RUNDELL, M. & RYCHLÝ, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX 2008*.

KILGARRIFF, A., RYCHLY, P., SMRZ, P. & TUGWELL, D. (2004). The Sketch Engine. In *Proceedings of EURALEX 2004*. Lorient, France, pp. 105–116.

KILGARRIFF, A. & TUGWELL, D. (2002). Sketching words. In Corréard, M.-H. (ed.) Lexicography and natural language processing: a festschrift in honour of BTS Atkins. Euralex, pp. 125–137.

KLOSA, A. (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher. In Kallmeyer, W. & Zifonun, G. (eds.) Sprachkorpora. Datenmengen und Erkenntnisfortschritt, Jahrbuch des Institut für Deutsche Sprache. Walter de Gruyter, pp. 105–122.

KRISHNAMURTHY, R. (2002). The Corpus Revolution in EFL Dictionaries. In *Kernerman Dictionary News*, 10.

LETURIA, I. (2014). *The Web as a Corpus of Basque* (PhD Thesis, UPV/EHU).

LINDEMANN, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. In *Procedia - Social and Behavioral Sciences*, 95, 249–257.

LINDEMANN, D. & San Vicente, I. (2015). Euskarazko maiztasun lemategia gaurko teknologien ikuspuntutik. In Fernández Fernández, B. & Salaburu Etxeberria, P. (eds.) Ibon Sarasola, gorazarre. Homenatge, homenaje. Bilbo: UPV/EHU, pp. 441–456.

LINDEMANN, D., Saralegi, X., San Vicente, I., Manterola, I. & Nazar, R. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density language pair. In *Proceedings of the XVI Euralex International Congress*. EURALEX, Bolzano: EURAC, pp. 563–576.

MITXELENA, K. & SARASOLA, I. (1988). *Diccionario general vasco - Orotariko euskal hiztegia*. Euskaltzaindia; Editorial Desclée de Brouwer.

NIDA, E.A. & TABER, C.R. (2003). *The Theory and Practice of Translation*. 4th ed. Brill.

RESNIK, P., OLSEN, M.B. & Diab, M. (1999). The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues." In *Computers and the Humanities*, 33, 129–153.

RUNDELL, M. & STOCK, P. (1992). The corpus revolution. In *English Today*, 8, 45–51.

SANZ, Z., ZUBILLAGA, N. & URIBARRI, I. (2015). Estudio basado en corpus de las traducciones del alemán al vasco. In Sánchez Nieto, M.T. (ed.) Corpus-based Translation and Interpreting Studies: from description to application. Berlin: Frank & Timme, pp. 211–235.

SARASOLA, I. (1982). *Gaurko euskara idatziaren maiztasun-hiztegia: 1977ko corpus batean oinarritua*. Donostia: Gipuzkoako Aurrezki Kutxa Probintziala.

SCHMID, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, pp. 47–50.

DE SCHRYVER, G.-M., JOFFE, D., JOFFE, P. & HILLEWAERT, S. (2010). Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. In *Lexikos*, 16.

SINCLAIR, J. (2005). Corpus and text-basic principles. In Wynne, M. (ed.) Developing linguistic corpora: A guide to good practice. Oxford: Oxbow Books, pp. 1–16.

SVENSÉN, B. (2009). *A handbook of lexicography: the theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

TEUBERT, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In Altenberg, B. & Granger, S. (eds.) Lexis in Contrast: Corpus-Based Approaches. John Benjamins Publishing, pp. 189–214.

TIEDEMANN, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 2214–2218.

UZEI. (2004). *Maiztasun Hiztegia*. Donostia: UZEI.

WOLFER, S., KOPLENIG, A., MEYER, P. & MÜLLER-SPITZER, C. (2014). Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses. In *Proceedings of the XVI Euralex International Congress*. Bolzano/Bozen: Eurac, pp. 281–290.

Resumen

La lingüística de corpus ha revolucionado la lexicografía: los datos de frecuencia, los ejemplos de uso y las traducciones son, entre otros, los recursos que nos ofrecen los extensos corpus de texto. La frecuencia de uso de una palabra y la frecuencia de visitas de diccionario de la palabra en cuestión están en relación; además, estos datos se pueden integrar en la entrada misma de diccionario. Por lo tanto, es interesante tomar listas de frecuencia como punto de partida para definir un lemario de diccionario.

En la lexicografía bilingüe se pueden utilizar los datos extraídos de corpus paralelos. En los últimos años se han creado varios corpus paralelos con euskera, como por ejemplo el corpus literario alemán-euskera. De estos corpus paralelos podemos extraer ejemplos de uso y sus respectivas traducciones, los que pueden resultar de ayuda en el trabajo lexicográfico, así como de información para el usuario. Esa última opción se está difundiendo en la actualidad a través de algunos portales de diccionario.

Résumé

La linguistique de corpus a constitué une véritable révolution dans le monde de la lexicographie, en effet, les larges corpus de texte offrent de multiples ressources telles que les données de fréquence, les exemples d'utilisation et les traductions. La fréquence d'utilisation d'un terme et la fréquence de visites de dictionnaire du terme en question sont en étroite relation ; en outre, ces données peuvent être intégrées dans l'entrée même de dictionnaire. Il est, par conséquent, très intéressant de prendre les listes de fréquence comme point de départ pour définir une liste d'entrées de dictionnaire.

Dans la lexicographie bilingue on peut utiliser des données extraites de corpus parallèles. Ces dernières années plusieurs corpus parallèles comprenant l'euskara ont été créés, comme par exemple le corpus littéraire allemand-euskara. Nous pouvons extraire de ces corpus parallèles des exemples d'utilisation et leurs traductions respectives, qui peuvent être très utiles dans un travail lexicographique, tout en apportant une information précieuse à l'utilisateur. Actuellement, cette dernière possibilité est en train de se répandre grâce à certains sites de dictionnaire.

Abstract

Corpus linguistics have revolutionized lexicography: data on frequency, usage examples, and translations are, among other things, the resources offered by extensive text corpora. The frequency of use of a word and the frequency of dictionary visits to that word are related; furthermore, these data can be integrated into the dictionary entry. Therefore, it is interesting to take frequency lists as a point of departure in defining the entry list of a dictionary.

Data obtained from parallel corpora can be used in bilingual lexicography. In recent years, a number of parallel corpora have been created with Basque, such as the literary German-Basque corpus. From these parallel corpora we can extract examples of use and their respective translations, which can be helpful not only in lexicographic work, but also in providing information to the user, a service that is becoming more and more common through the use of dictionary websites.