6

Baliabide lexikoen sarea: Baldintza filologiko eta tekniko zenbait

DAVID LINDEMANN
Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

IÑAKI SAN VICENTE
Elhuyar

1. Sarrera

Humanitate Digitalen (HD) arloan, berrikuntza garrantzitsuak izan dira azken urteotan, eta informatikarekiko lankidetza dela medio, metodologia arlo konputazionaletara zabaldu egin du filologiak. Testu klasikoen edizio digitalak burutu eta prozesamendu konputazionalerako prestatzea da Humanitate Digitalen ardatz nagusi bat. Argitalpen historikoen edukiak testu digital bihurtzeaz gain, edizio digitala metadatuez osatzera igaro dira HD arloko ikerlariak, hala nola, anotazio filologikoak (esaterako, forma historikoa egungo formarekin lotzen dituztenak), edo anotazio linguistikoak, adibidez, hitz bati dagokion lema, analisi morfosintaktikoa edota semantikaren inguruko anotazioak. Bestetik, errekurtsoen artean loturak daramatzaten poligrafiak sortzeko ahalmena ematen du edizio digitalak (Smith, 2004).[1]

Hiztegigintza historikoan ere metodologia konputazionalen etorreraren lekukoak gara, Europako hainbat hizkuntza-komunitatetan burutu diren ikerlanei erreparatzen badiegu. Hiztegi historikoa banakako baliabide itxi bezala ulertu beharrean, haren edukiak beste hainbat baliabidetako edukiekin lotzeak abantaila nabarmenak ekartzen dizkio edizioaren erabiltzaileari. Hala egitea ez da batere ideia berria, baina, Orotariko Euskal Hiztegiaren (OEH, Mitxelena & Sarasola, 1988) ekoizpen prozesuan egin zen moduan, hiztegi historikoen edukiak kartoizko txarteletan idatzi eta indexatu beharrean, hiztegi sarrerak eta haien metadatuak errepresentatzeko eredu digital aurreratuak ditugu egun. Eta haiek erabilita, edizioaren egileak eta erabiltzaileak biek dituzte onurak. Edizioen elkarreragingarritasuna bermatzeko asmotan, estandarrak garatzen ari dira. Artikulu honetan, joera berri hauen oinarri batzuk ikusi eta metodologia konputazionalaren zenbait aplikazio ikusiko ditugu, baita nazioarteko arloaren egoera euskararen eremura ekartzeko zenbait baldintza identifikatu ere. Oso oinarrizko esperimentu bi aurkeztuko dugu, baliabide lexikal ezberdinen edukiak elkartzeak zer aukera zabaltzen duen irudikatzeko.

[1] Cologne Centre for eHumanities erakundeak burututako *Vedaweb* proiektua aipa daiteke edizio digital aurreratuaren adibide gisa, Rig-Veda-ren edizio ezberdinak lerrokaturik eta metadatuekin batera interfaze publiko baten bitartez eskaintzen dituena, ikus https://vedaweb.uni-koeln.de/. Kopenhageneko ONP *Old Norse Prose*, berriz, baliabide ezberdinak interfaze bakarretik atzitzeko eskaintza da, ikus http://onp.ku.dk/onp/onp.php

2. Hiztegien sareak

Euskarazko hiztegi guztiak, klasikoak nahiz egungoak, banakako *stand-alone* baliabide gisa ditugu eskura, elkarrekiko lotura esplizituritk gabe. Hiztegien ohiko egituran eta ekoizpenerako ohiko lan-fluxuan datza atomizazioaren zergatia: lemategia zehaztu ondoren, lema bakoitza argibide lexikografikoez hornitzera igaro ohi da hiztegigilea. Lemaren azpian, desanbiguazio sintaktiko edota semantikoa txertatzen du, hiztegi-sarrerari zer egitura eman nahi dionaren arabera. Aldi berean, edukien aurkibide bakarra lemategia izan ohi da, kasu batzuetan, sarrera batetik beste sarrera batera bidaltzen duten zeharkako erreferentziez osaturik. Hiztegi batean hiztegi beretik kanporako erreferentziak jasotzen badira, OEH-k lemaren edo haren aldaeraren agerpena hiztegi aurrekarietan aipatzen duen bezala, kanpo-erreferentzia horiek ez dira, orokorrean, edizio digitaletan landu, hau da, ez dago aplikazio informatiko batek interpreta lezakeen lotura esplizituritk aipuaren jatorrizko agertokira edo sarrera horrek beste kanpo baliabideetan izan dezakeen informazio osagarrietara.

Hori dela eta, lema-ikurra (hots, lema ordezkatzen duen hizki segida) baino ezin izan da ustiatu baliabideetan zehar loturak eraikitzeko. *Euskalbar* bilaketa-tresna[2] eta *Elhuyar Hiztegi App*[3] horrelako ekimenak dira, bilaketa bakarra hainbat hiztegitan aldi berean egitea ahalbidetzen dutenak. Bilatzen den hizki-segida bat euskarazko hainbat online-hiztegitako sarbidetara bidaltzen dute tresna horiek, emaitzarik izango ote den aurretik argitu edo bilaketa lema-ikurraz bestelako edukietara murriztu edo hedatu ahal izan gabe. Bilatzen den hizki-segida bera hainbat hiztegitako bilaketa-eremuetan eskuz sartzeko (eta emaitzak ikustarazteko) prozesua bizkortzea baino ezin diote erabiltzaileari eskaini, beraz, aipatutako tresnek. Murriztasun horren zergatia hiztegi elektroniko gehienen datu-base egituran datza: lema-ikurrak baino ez ohi dira bilaketak ahalbidetzeko indexatu, ezta datu-baseon edukiak API[4] bidez zuzenean atzitzeko prestatu ere.

Gauzak horrela, *hiztegi sare* motako baliabidea ezin izan da ekoitzi euskararako, hau da, hainbat hiztegitako edukiak ezin izan dira batu eta lema-ikurra loturak antolatzeko pibot elementutzat harturik elkarrekin eskura jarri, nazioartean hainbat kasutan egin den legez. Alemanezko *Wörterbuchnetz* hiztegien bilduma,[5] galegozko *Dicionario de Dicionarios*,[6] portugesezko *Corpus Lexicográfico do Português*[7] eta nederlanderaren *Geïntegreerde Taalbank*[8] gertuko adibideak dira.[9]

Nederlanderaren *Geïntegreerde Taalbank* baliabidea Europa mailan aurrekaria dela esan daiteke. Aipatutako beste ekimenetan bezalaxe, hiztegi historikoen edukiak bateratu egin dituzte, baina besteetan ez bezala, lema-ikur historikoak egungo nederlanderazko formekiko lotura esplizituez hornitu dituzte; horrela, lema-ikur historikoak nahiz egungoak erabil daitezke bilaketak egin eta beti forma historiko zein egungoari lotutako hiztegi-edukiak ikustarazteko. Jokabide hori OEHk euskararako erakusten duen irizpide filologikoarekin bat dator, hau da, forma historikoak

[2] Ikus https://github.com/euskalbar/euskalbar

[3] Ikus https://www.elhuyar.eus/eu/site/prentsa-aretoa/147/hiztegiapp-mugikorretarako-aplikazio-berria

[4] Ingelesez, *Application programming Interface*, aplikazio edo plataforma batek batek bere edukiak hirugarrengo baten eskura jartzeko erabiltzen den zerbitzua. Eduki horiek programa bezero baten bidez atzitzen dira.

[5] Ikus http://www.woerterbuchnetz.de

[6] Ikus http://sli.uvigo.es/DdD/

[7] Ikus http://clp.dlc.ua.pt/DICIweb/

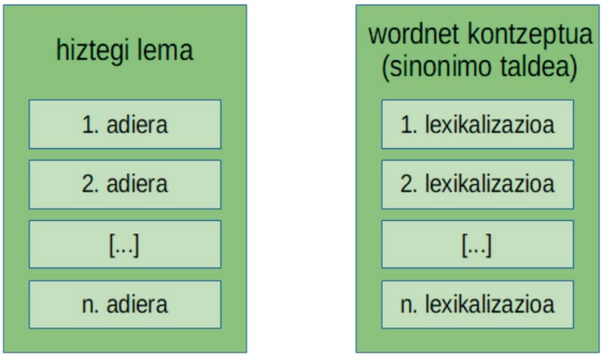
[8] Ikus http://gtb.inl.nl/search/

[9] Artikulu honetan arazo filologiko eta teknikoak ditugu hizpide. Alabaina, aipatu beharra dago hiztegien egile-eskubideak ere oztopoa izaten dira hiztegien edukiak berrerabiltzeko.Egile-eskubideak argitalpena baino 70 urte beranduago desagertzeak ahalbidetzen du hiztegi historikoen edukiak berrerabiltzeko; hiztegi berriagoen kasuan, berrerabiltzeko hitzarmenak behar dira, lizentzia irekiko baliabideak ez badira.

egungo formarekin lotzea, OEHk hiztegi historikoetako lema-ikurrak egungo lema-ikurren azpian erreferentzia gisa zerrendatzen dituen gisan. Baina nederlanderazko baliabidean, egungo lemak ez ezik, forma historikoak ere bilaketarako indexatu dituzte, horrela aurkibidea corpus lexikografikoan agerpena duten lemen aldaera guztietara hedatuz.

Esan bezala, lema-ikurraz bestelako edukiak ezin izan dira elkarrekiko loturaz hornitu, eta horrela gertatzen da aipatu ditugun nazioarteko hiztegi sareetan. Egoera hori gaindi dezakeen Hiztegi Matrix izenda litekeen baliabideak[10] lema-ikurrez gain adierak ere hartuko lituzke aintzat hiztegietan zehar loturak ezartzeko. Adierak elkarrekin lotzea (*sense linking* deritzona) egun pil-pilean dagoen erronka dela esan daiteke. Adierak (*word senses*) lema homografo eta polisemikoen kasuan desanbiguatzea Lengoaia Naturalaren Prozesamenduaren (LNP) erronka nagusietako bat da. Azken urteetan hitzen errepresentazio bektorial berriei esker (*word embeddings*) aurrerapauso garrantzitsuak eman dira (Raganato, Camacho-Collados & Navigli, 2017). Hala ere, emaitzarik onenak lortzen dituzten sistemak gainbegiratuak dira, eta adierak eskuz etiketatuta dituzten datu bildumak behar dituzte desanbiguazioa egiten ikasteko. Tamalez, baliabide horiek oso urriak dira eta gainera kostu handia du horiek sortzea (Camacho-Collados & Pilehvar, 2018). Jakintza lexikografiko preziatua paperezko argitalpenetatik atera eta ustiapen konputazionalerako prestatzea, beraz, hutsune haiek betetzeko jokabide baliotsua izan daiteke.

Polisemiaren arazoaren aurrean, eta hiztegigintzari dagokionez, lema-ikurretatik abiatu beharrean kontzeptuetan oinarritzen diren baliabideak sortzea izan da erantzun bat. *Wordnet* bezalako baliabide lexikoek, esaterako, sinonimo-talde banaren bitartez lexikalizatzen diren kontzeptuak biltzen dituzte, hiztegietan adierak lemaren azpian biltzen diren bitartean (ikus 1 irudia). Kontzeptua sinonimo-taldeko kideen (hots, kontzeptuaren lexikalizazioen) adieretako bat da, definizio batez deskriba daitekeena eta beste kontzeptuekiko erlazio semantikoak dituena. Sinonimiak, eredu honen arabera, kontzeptu bera denotatzea esan nahi du, dagokion unitate lexikoaren erregistroari, dialektoari edo datazioari, adibidez, erreparatu gabe. Xehetasun horiek unitate lexikoari lotuta daude ikuspegi horretatik, eta ez kontzeptuari. Hiztegiak adiera mailan elkarrekin lotu ahal izateko jokabidea ematen digu Wordnet ereduak, beste hizkuntzetako Wordnet-etara nahiz bestelako baliabide lexiko-semantikotara,[11] kontzeptutik abiatzen diren loturek adierak zuzenean lotu, eta adieren desanbiguazioaren arazoa gainditzen baitute.

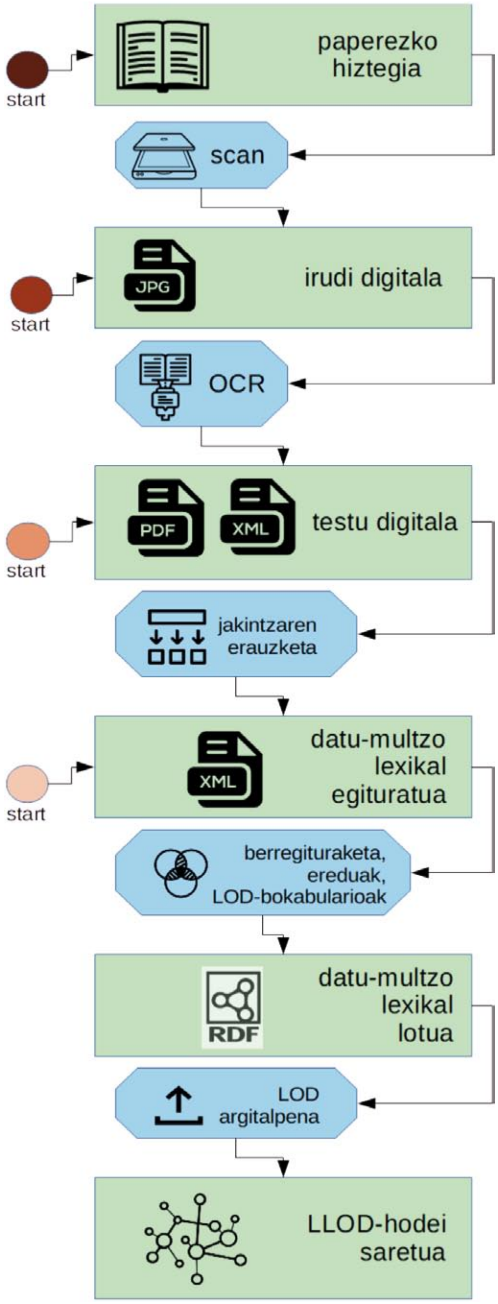


1. IRUDIA

Hiztegi lema eta WordNet kontzeptua

[10] *Elexis Lexicographic Infrastructure* ikerketa-azpiegitura proiektu europarrean (2019ean abiatuta, ikus http://elex.is) lema eta adiera mailan lotutako hiztegien bilduma eleanitza sortu nahi dute, *Dictionary Matrix* izenpean.

[11] Adibidez, danierazko hainbat baliabide lexiko adiera mailan lotu dituzte, danierazko Wordnet kontzeptuak pibot elementutzat harturik (Pedersen, Nimb, Olsen & Sørensen, 2018).



2. IRUDIA

Hiztegiak digitalizatzeko lan-fluxu orokorra

3. Hiztegiak saretzeko lan-fluxua

Hiztegien edizio digitalak burutzeko bide berriak jorratu dira azken hamarkadatan. Bibliotekonomian oro har gertatzen den bezalaxe, hiztegigintza historikoan ere egungo irizpideen araberako edizio digitalak, jatorrizko edizioaren edukiaren babesa bermatzeaz gain, edukiak makinak uler dezakeen formatuan errepresentatzea du helburu, hau da, testu-geruza duen faksimile digitala eskaintzeaz gain, hiztegiaren makroegiturari (lemategiari) eta mikroegiturari (sarreren barruko antolamenduari) dagozkion anotazioak gehitzea, makinak testu-zatia bakoitza item lexikografiko zehatz gisa interpretatu ahal izateko. Bestela esanda, testu-geruza hutsari anotazio semantikoak dituen metadatu-geruza bat gehitzen zaio. Errekurtsoaren barruko bilaketa aurreratuak ahalbidetzeko eta baliabide barruko nahiz kanporako lotura esplizituak esleitzeko baldintza teknikoak betetzen dira horrela. Nazioartean, hainbat hiztegi klasikotako edukiak horrela prozesatu izan dira, eta egun *machine readable* formatuan ditugu eskura.[12]

Hiztegi historikoen kasu askotan, Larramendi (1745) euskarazko klasiko ezagunaren kasuan, esaterako, digitalizazioa faksimile-irudiak ekoiztean gelditu egin da, eta, ondorioz, pixel-irudi haiei begira dagoen erabiltzailek ikus ditzakeen testu-edukiak ezin dira ordenagailuz lagunduta aurkitu ezta bestela prozesatu ere. OCR (karaktereen ezagutza optikoa) izango litzateke testu-edukiak azaleratzeko beharrezkoa den lan-urratsa. Hiztegi berriago gehienen kasuan, testu-geruza duen PDF formatuko argitalpena edota argitalpenaren oinarria izan den testu-fitxategia dugu, eta zenbait kasutan, are gehiago, mikroegitura lexikografikoa islatzen duen markaketa semantikoaz osatua da hiztegiaren testua; hala nola, Euskaltzaindiaren 2010ko Hiztegi Batuaren XML bertsioa.[13] Horregatik, eta 2 irudian azaltzen dugunez, hiztegi bat datu-multzo lexikal egituratu bihurtzeko lan-fluxuak abiapuntu ezberdina izan ahal du («start»), jatorrizko formatuaren arabera.

Hiztegiaren testu-geruza metadatu lexikografikoez osatzeko, ordenagailuz lagundutako metodologiak ere urrats garrantzitsuak eman dituela esan daiteke. «Lema-ikurra», «adiera», «definizioa» edo «itzulpen-ordaina» bezalako mikroegitura-elementuak atzeman eta markatzeko, ikasketa automatikoko algoritmoak darabiltzan tresna batekin esperimentuak egin ditugu, lizentzia libreko OEH-ren PDF bertsioa adibide hartuta.[14] Zenbait orrialdetan markaketa eskuz egin eta tresnak entrenatzeko erabili ondoren, hiztegi-sarrera bakoitzaren hasiera eta amaiera markatzeko gai izan gara hiztegi osoan zehar, sarrera oso laburra izan edo sarrera batek orrialde bat baino gehiago hartzen badu ere. Ikasketa Automatikoko algoritmoak malgutasuna erakutsi du eginkizun horretan, erregeletan oinarritutako sistema batek nekez izango lukeena. Sarreren hasiera eta amaiera zehazturik, sarrera bakoitzaren barrukoak zatitu eta etiketatzea izango litzateke hurrengo lan-urratsa, handitik txikira aurreratuz («cascaded approach», ik. (Khemakhem, Foppiano & Romary, 2017)). Tresna entrenatzeko datuak eskuz prestatu behar dira, zatiketa automatikoaren emaitzen (eskuzko) ebaluazioak doitasun maila egokia lortu arte.

Edizio digitalen elkarreragingarritasunaren bila, formatu-estandarra sortzeko ahaleginak ere bultzatu dira. Hiztegien barruko antolamendua errepresentatu eta, loturak ezartzeko asmotan, estandar baterantz moldatzeko ekimenen artean, TEI-XML ingurukoak aipatu behar ditugu,[15] hiztegi digitalen egiturak eta markaketa-lengoaiak berdintzea helburu dutenak. Hiztegiak eta bestelako baliabide lexiko-semantikoak elkartzeko, berriz, azken urteotan maiz aipaturiko estandarrak sortzen ari dira *Linguistic Linked Open Data* eremua lantzen duen komunitatean, haren baitan dagoen *Ontolex* W3C lan-taldean,[16] bereziki. Web semantikoaren RDF estandarra (Berners-Lee, 2006) erabili eta hiztegien edukiak eredu formal batean jasotzeko ahaleginak dira *Ontolex-core* («lemon») eta *Ontolex-lexicog* ereduak. Ereduok eskaintzen dituzten klaseak elkarrekin lotuta

[12] Digitalizazio lan-fluxuen inguruan, ikus, adibidez, https://digilex.hypotheses.org/ blogean baturiko txostenak. Oraingoz lema mailan baino elkarrekin loturarik ez duten hiztegi historikoen bilduma bat ikus http://woerterbuchnetz.de helbidean.

[13] http://www.euskaltzaindia.eus/dok/eaeb/hiztegibatua/ helbidean eskuragarri.

[14] Ikus https://digilex.hypotheses.org/250

[15] Ikus https://github.com/DARIAH-ERIC/lexicalresources

[16] Ikus https://www.w3.org/community/ontolex/

daude, zehazki definituriko propietateen bitartez, eta haien arabera sailkaturiko baliabide lexikalen edukiak aurkigarriak izango dira mikroegitura-elementuak islatzen dituzten klaseak eta haiekin lotutako propietateak zehazturik. Horretaz gain, *Ontolex-core* ereduko klaseek *Wikidata* baliabidearen zenbait klaserekiko loturak eskaintzen dituzte.[17] Wikidata baliabidearen barruan, berriz, milioika kontzeptu daude bilduta, hiztegietako adierekin lotu daitezkeenak. Kontzeptuak biltzen dituen *Wikipedia* ez bezala, *Wikidata* datu lexikalak ere batzen hasi da. Osagai horiekin, *Wordnet* bezalako egitura bat eraiki daiteke *Wikidata*-n bertan, eta *Wikidata* bera izan liteke, beraz, Hiztegi Matrixaren ardatz eta kokapena. Nabarmentzekoa da Euskararen pozisionamendua *Wikidata*-ko datu lexikalei dagokienean, Euskal Wikilarien Elkartearen eta Elhuyar fundazioaren elkarlanari esker.[18]

Adiera mailan lotutako Hiztegi Matrix delakoa etorkizuneko erronka bada ere, laburbiltzeko, bi baldintza nagusi bete behar dira Hiztegi Matrix-aren parte izan daitezkeen datu-multzoak sortzeko; lehenik, datu lexikalak egituratuak izan behar dute, XML edo RDF bezalako ereduak erabilita. Hiztegi-sarrera baten barruan, gutxienez honako mikroegitura-elementu hauek markaketa eraman behar dute, hiztegia sareratu ahal izateko: lema-ikurra, kategoria eta adiera.[19] Bigarren baldintza litzateke loturak ezartzea: Alde batetik, lema-ikur hutsaz haratago, kategoria duen lema bakoitzari (*lempos entity* bakoitza) eta, bestetik, adiera bakoitzari identifikatzailea esleitzea, identifikatzaile hori propietate sorta batez kanpoko baliabideetako beste identifikatzaile batzuekin lotu ahal izateko. Lema mailako loturak ezartzeko bide nahiko zuzena dagoen bitartean, lema ikurra bera horretarako balia daitekelako, sense linking deritzona askoz ere eginkizun konplikatuagoa da. Joera berrienek hemen ere ikasketa automatikoan oinarritutako teknikak erabiltzen dituzte, eskuzko lana errazteko asmotan (ikus eztabaida Saurí, Mahon, Russo & Bitinis, 2019 lanean). Erabateko doitasunari hurbiltzeko, alabaina, hiztegigilearen eskuzko lanak ezinbestekoa izaten jarraitzen du, eta algoritmoen garapenean ez ezik, eskuzko lanean ere inbertitzen jarraitzen dute era masiboan arloko egungo proiektuek. Polonierazko Wordnet baliabidea hedatu eta haren kontzeptuak ingelesezko Wordnet-ekoekin lotzeko, esaterako, dozenaka hiztegigilek dihardute polonieraren inoizko baliabide lexiko-semantiko handiena sortu eta mantentzeko ahaleginean (Maziarz, Piasecki, Rudnicka, Szpakowicz & Kedzia, 2016).

4. Bestelako baliabide lexikalak

Hiztegietan zehar ez ezik, lemak eta adierak dituen bestelako baliabide lexikaletan zehar ere loturak ezartzea ezinbesteko baldintza da hainbat aplikazio burutzeko. LNP alorrean kokatu ohi diren baliabide lexiko-semantikoetatik bat aipatu dugu goian: *Wordnet* kontzeptusarea (Miller, Beckwith, Fellbaum, Gross & Miller, 1990). Horretaz gain, testu-corpusetatik erauzitako maiztasun lemategiak eta hizkuntzalaritza konputazionaleko lexikoiak ere lotu ditugu lema-ikur mailan elkarrekin eta hiztegietako lemategiekin. Hiztegi elebidun berri bat sortzeko, euskarazko oinarrizko lemategia sortzea baitzen gure asmoa (EusLemStd, Lindemann & San Vicente, 2015).

[17] Ikus https://www.wikidata.org/wiki/Wikidata:Lexicographical_data eta https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/RDF_mapping

[18] Ikus euskarazko lexemen kopurua *Wikidatan* https://tools.wmflabs.org/ordia/language/ helbidean, eta adibide bat hemen: https://www.wikidata.org/wiki/Lexeme:L48738. Wikidata-ko euskarazko lema guztiak honako SPARQL-query honen bitartez ikus daitezke: https://w.wiki/Byt

[19] Mikroegitura-elementu gehiago markatuz gero, haien edukia ere aurkigarri eta loturak ezartzeko gai bihurtzen dira, hala nola, aipuak, *Geïntegreerde Taalbank* gertatzen den moduan.

Horretarako, gaurko corpus handienetan, hau da, ETC Egungo Testuen Corpus (Sarasola, Landa & Salaburu, 2013) eta Elhuyar 200M-Webcorpusean (Elh200, Leturia, 2014), eta, aldi berean, hiztegi nagusietariko batean[20] agerpena duten lema-ikurrak zehaztu ditugu. Horrela, hiru baldintza betetzen zituzten lema-ikurren zerrenda lortu dugu. Lehenik, lema-ikur historikoak baztertu ditugu, hau da, gaurko erabilera islatzen duten corpusetako batean agerpen-atalasea[21] gainditzen ez dutenak. Bigarrenik, hiztegi nagusietan inon jaso ez direnak baztertu ditugu, corpusetan agertzen den guztiaren artean jasotzeko mereziko luketeen neologismoak iragaztea ez baita hiztegigintza elebidunaren zeregina, Lexikoaren Behatokia (Artola, Ezeiza, Gurrutxaga, Sagarna & Urkia, 2017) bezalako ekimen berezituaren eginkizun erraldoia baizik. Eta hirugarrenik, hiztegien eta corpusen intersekzioa islatzen duen lemategia corpusetako maiztasun-datuekin hornitu dugu, horrela EusLemStd lemategia abiapuntua duen hiztegi berrian lemen maiztasunari buruz argibideak eman ahal izateko.


```xml
<homograph homograph="aditu" corpus_counts="42042">
    <ADI lemma="aditu" pos="ADI_SIN" corpus_counts="18989">
        <sense synset="30-00588888-v" equivs="understand"/>
        <sense synset="30-02169702-v" equivs="hear"/>
        <sense synset="30-02571901-v" equivs="heed mind listen"/>
    </ADI>
    <IZE lemma="aditu" pos="IZE_ARR" corpus_counts="13945">
        <sense synset="30-09617867-n" equivs="expert"/>
        <sense synset="30-10557854-n" equivs="scholar scholarly_person bookman"/>
    </IZE>
    <ADJ lemma= "aditu" pos="ADJ_ARR" corpus_counts="5486">
        <sense synset="30-02226162-a" equivs="adept expert skillful"/>
    </ADJ>
</homograph>
```


3. IRUDIA

'aditu' lema EDBL eta EusWN datuak eta corpus maiztasunak

Aipatutako lanean euskal Wordnet (EusWN, Pociello, 2007) eta EDBL (Aldezabal *et al.,* 2001)[22] ere aintzat hartu genituen, corpus eta hiztegiekiko intersekzioak definitzeko. Hutsetik sortzekoa litzatekeen hiztegi elebidun berria eduki gehiagoz osatzeko ere erabili genituen bi baliabide horiek (Lindemann & San Vicente, 2016). EusTagger tresnak (Aduriz *et al.,* 1996) corpuseko formei lema eta kategoria esleitzen dizkie, EDBL-eko datuetan oinarritua. EusLemStd zerrendako lemak, hau da, Elh200 corpusean gutxienez 20 aldiz eta, aldi berean, hiztegi nagusietako batean agertzen direnak, EusTagger-rek asoziaturiko hitz kategoriez anotatu genituen, bigarren pauso batean EusWN-eko adierak ere gehitzeko. EusLemStd sarreren heren bat estali ahal izan dugu horrela gutxienez kategoria eta adiera banarekin. 3 irudian erakusten dugu emaitzen adibide bat, aditu lema, lehenengo maila batean hitz kategoriaren arabera antolatuta, haren azpian *Open Multilingual Wordnet* (Bond & Foster, 2013) bitartez lortutako ingelesezko itzulpen-ordainak eta Elh200 corpuseko agerpen-kopuruak adierazten ditugula.

[20] Orotariko Euskal Hiztegia (Mitxelena & Sarasola, 1988), Hiztegi Batua (Euskaltzaindia, 2010), Euskal Hiztegia (Sarasola 1996) eta Elhuyar Euskara-Gaztelania hiztegiaren euskarazko lemategia (Azkarate, Kintana & Mendiguren, 2006) dira eskuragarri izan ditugun lemategiak.

[21] Hogei agerpenetan finkatu dugu atalasea, corpus hizkuntzalaritzako usadio bati jarraituz.

[22] Euskararen datu base lexikala (EDBL) corpusetan metadatu morfosintaktikoak gehitzeko eta zuzentzaile ortografikoetan erabiltzen da, besteak beste.

Horrela lortutako hiztegi-zirriborroa hiztegi elebidunaren abiapuntua balitz, eta eduki guztiak irudian dugun bezalako markaketaz hornituko balira, zirriborroa erauzteko baliatu zituen iturriak aberasteko lanean jardungo luke hiztegigileak aldi berean, eskuzko lanaren balorea bikoiztuz, elkarreragingarritasun iraunkorreko *bootstrappingloop* deitu daitekeen lan-fluxu batean. Corpus maiztasun atalasea gainditu eta EusTagger-rek kategoria esleitu arren EusWN adierarik ez duten EusLemStd zerrendako lemak izango lirateke landu beharrekoak. Eta lema hauek hiztegigintza elebidunean lantzeak haien adierak zehaztea esan nahi duenez, EusWN hedatzeko ekarpena dakar. Bestetik, lema batek dituen adiera guztiak EusWN baliabidean islatuta dituen edo ez galdetu beharko da; horretarako, hiztegietan duen adiera-kopuruari begiratzea izan daiteke arazoari hurbiltzeko bide bat (Lindemann & San Vicente, 2016).

Euskarazko hiztegiek ez dutela baliabide bakoitzetik kanpora bidaltzen duen lotura explizituri esan dugu. Corpusetako adibideak hiztegi sarrerei edo adieretako bakoitzari lotuta agertzen badira, hiztegiaren parte gisa jaso ditu hiztegigileak, Egungo Euskararen Hiztegian (EEH, Sarasola, 2008), esaterako, gertatzen den moduan. Hiztegitik kanpora lotura esplizituri ezean, erabiltzaileak baliabide ezberdinen artean salto egiteko moduak badaude, hala ere. Elhuyar hiztegi elebiduneko erabiltzaile-interfazeak[23] eskaintzen dituen aukerak adibide moduan azalduko ditugu.

Batetik, hiztegi horretako sarrera elebidunetan, corpus paraleloan bilaketa egiteko aukera ematen da. Bilaketa egiteko bi hizkuntzetako lemak erabiltzen dira, eta horrela bilaketa bi ordainak agertzen diren esaldi-bikoteetara mugatu. Adieren desanbiguazioa, beraz, ordainaren bidez egiten da, bilaketa egiten den momentuan bertan. Jokabide horren abantaila begi-bistakoa da: corpusetako adibideak ez dira euskarazko lemaren beste adierei dagozkienekin nahasten. Aldi berean, horrek adibideak hiztegian agertzen diren ordainetara mugatzen ditu, bestelako ordainak bilatzeko corpusera zuzenean jo beharko genukeelarik.

Bestetik, Elhuyarrek bere hiztegi tekniko eta terminologikoak elkar-lotuta ditu. Horri esker, erabiltzaileak hiztegi-sarrera bat kontsultatzen duenean aukera du sarrera hori aipatutako kanpo baliabideetan kontsultatzeko. Horren muga, berriz ere, erakunde ezberdinek garatutako baliabideak elkarlotzeko aukeran datza. Hori bilaketa bidez ahalbidetu liteke (*Euskalbar* tresnak egiten duena), baina ez hiztegi-sarrera desanbiguatuen mailan.

5. Sarearen aplikazioak

3. atalean aurkeztutako lan-fluxuaren emaitza datu-multzo lexikalen sorta da, eta haien elkarketa lema-ikurren edota adieren bitartez baliabide lexikalen sarea, edo, neerlanderazko izendapenereduari jarraituz, hizkuntza-banku integratua. Horrelako sarearen aplikazioen artean, erabiltzaile-interfaze baten bitartez bilaketa aurreratuak eskaintzea aipa daiteke lehendabizi, *Geïntegreerde Taalbank*-ek duen interfaze publikoak ahalbidetzen duen bezala. Galizierazko baliabide lexikalen sareak *Recursos Integrados da Lingua Galega* du izena,[24] idatzizko lekukotza eta lema-ikurraren araberako bilaketak ahalbidetzen ditu hainbat baliabidetan aldi berean, hiztegietan, GalNet *Wordnet*-ean nahiz testu-corpusetan, lemaren kategoria gramatikala zehazturik.

[23] http://hiztegiak.elhuyar.eus helbidean eskuragarri.

[24] Ikus http://sli.uvigo.es/RILG

Bigarrenik, sare bihurtutako datu lexikalen bilduma horrelako beste sare batzuekin elkar daiteke, egiturak TEI-XML edo *Ontolex*-RDF bezalako nazioarteko estandar bati jarraitzen badio. Lema mailan elkarturik dauden bilduma batzuk goian aipatu ditugu. Bestetik, *Wordnet* sare semantikoaren ereduari jarraitzen dioten bildumak aipagarriak dira, azken urteotan maiz aipaturiko *BabelNet* ekimena tartean (Navigli & Ponzetto, 2010),[25] *Open Multilingual Wordnet* eta *Wikipedia* baliabideen edukiak, besteak beste, kontzeptu mailan batzen dituena, doitasun handiko algoritmo baten bitartez.[26] BabelNetek bere edukiak aberasten jarraituko du, eta hiztegi historikoen edukiak *BabelNet*-eko kontzeptuekin lotzea da, hain zuzen ere, *Elexis* ekimenean jarri duten erronketako bat (Declerck, McCrae, Navigli, Zaytseva & Wissik, 2018).

Hirugarrenik, baliabide lexikalen sarea filologia arloko ikerketa kuantitatiboen iturri gisa erabil daiteke, edota hurbiletik aztertzeko laginak sortzeko. Euskarazko zenbait datu lexikal elkartu ditugu, baliabide ezberdinetatik erauzketak eginda. Jarraian, elkarketa horren aplikazioetarako bi adibide emango dugu.

5.1. *Larramendiren lema berrienerabilera gaur*[27]

Goian aipatu dugun euskarazko hiztegi klasikoa, Larramendirena, alegia, ez dugu oraindik edizio digitalean. Alabaina, digitalizazioa burutu ezean, bidezidorra dugu beste hiztegi baten edukiak erabiliz. Izan ere, Ibon Sarasolaren hiztegi elebakarretan '*1745' marka darama lemak, Larramendiren hiztegia lehenengo agerpen dokumentatua baldin bada. Bestetik, egungo lema-ikur baten azpian jasotzen du Sarasolak Larramendik erabilitako forma (OEH-n gertatzen den bezalaxe). Larramendiren 'beguiordeac' (s.v. 'anteojos'), adibidez, islatzeko, grafiari eta, pluraleko forma denez, morfologiari dagokionez gaurko lematizazio estandarrari erantzuten ez dion forma beharrean, 'begi-orde' bezala zerrendatzen dute Sarasolak eta OEH-k, eta begiordeak da Larramendiren hiztegiko aipuan OEH-n ematen duen forma. Egungo euskararen lematizaziorako egokitzapena ere badugu, beraz, Sarasola/OEH hiztegietan. OEH-tik lemaz bestelako eremuetara murriztutako bilaketak egun ezin ditugu ordenagailuz lagunduta egin. Alabaina, OEH-ko begi-orde adibideak prototipikoki erakusten du lematizazio historikoa interpretatzeko jokabidea ('beguiordeac'>'begiordeak'>'begi-orde'), eta hain zuzen Sarasolak eta OEH lantzen duen taldeak egindako egokitzapen horregatik izango gara gai Larramendiren lemak gaurko euskararen beste baliabide batzuetan bilatzeko.

Euskal Hiztegiaren 1996ko edizioa XML bihurtu dutenez (Arriola *et al.,* 2003), hiztegi osoan zehar aurki eta erauz ditzakegu edukiak, mikroegitura-elementua zehazturik, datazioarena, esaterako. Larramendiren lanean lehendabiziko agerpena duten lemak 3.420 '*1745' marka daramaten horiek direla onartuz gero, eta lema haien zerrenda egungo corpusetatik erauzitako maiztasun-zerrendekin erkatzen badugu, Larramendiren lema berrien arrakastaz zerbait esan dezakegu, eskuzko inolako bilaketarik egin gabe.

[25] Ikus http://babelnet.org

[26] BabelNet baliabidetik euskaraz-ingelesezko hiztegi elebiduna erauzteko gure esperimentu eta ebaluaketaz ikus (Lindemann & Kliche, 2017). EusLemStd zerrendaren estaldura nabarmen handiagoa eskaini du BabelNet-ek (%40), WordNet soilik erabiltzearen aldean (%31), doitasuna maila berean mantentzen duen bitartean (%85).

[27] Azkarate & Lindemann (2018) lanean antzeko ariketa bat aurkeztu dugu, Sabindarren neologismoen gaurko erabileraren inguruan.

1 eranskinean '*1745' lemen maiztasun-zerrenda ematen dugu, ETC eta Elh200 corpusetik erauzitako maiztasun-zerrendan dituen tokiak adierazita. Gaurko euskararen 750 gehien erabilitako lemetatik 40 Larramendik lehen aldiz jasotakoak direla izango litzateke ondorioa, ereduzko testuetan nahiz internetean topaturiko euskarazko edukietan. Oro har, 3.420 lema-ikur haien erdiak (1.742) Elh200 corpusean hogei agerpen edo gehiago dutenez, gaurko erabilera frogatutzat eman daiteke, beste erdiaren %44 (1678tik 733) hogei aldiz baino gutxiago agertzen den bitartean; gainontzeko 945 lema ez dira behin ere agertzen.

5.2. *Twitter corpusaren azterketa*

Azkenik, sare sozialetara begira jarri nahi izan dugu. Horretarako Twitterretik erauzitako 40 Miloi hitzeko corpusa erabili dugu. Corpus orokorra da, 2017 urtetik 2019ko iraila bitartean bildua, eta hizkuntza automatikoki detektatuta. Twitterren erabiltzen den hizkuntza aztertzeak berez erronkak ditu. Izan ere, Twitterren erabiltzen den hizkuntza askotan kaleko ahozko hizkeratik hurbilago dago idatzizko hizkuntzatik baino. Ohiko testuetan topatzen ez ditugu zenbait fenomeno tratatu behar dira. Adibidez, gaztelania eta euskara nahasten dituzten esaldiak (Code Switching izenez ezagutzen dena, adib. 'Ta atzo kede fatal!!!', 'Ingleseko azterketa de puto kulo itezunen...'), enfasia adierazteko hitzak maiuskulaz idaztea ('A ze pelikula TXARRA ikusi dugun gaur') ala bokalak errepikatzea (adib., 'ooona!'), edo lekua aurrezteko hitzak moztea (adib., 'msdz' → 'mesedez', 'mlskr' → 'milesker', 'pxkt' → 'pixka bat'). Fenomeno horiek denek noski prozesamendu automatikoa zailtzen dute, harik eta berariazko aurreprozesatze bat eskatzen dutelarik (Alegria *et al.*, 2015). Aurreprozesatze hori bereziki da beharrezkoa hiztegigintza arloko aplikazioetarako, edo lemak beste baliabideetako loturez hornitzeko, forma ez estandarren lemak identifikatu nahi baditugu.

Sare sozialetako datuek erronkak planteatzen badituzte ere, abantailak ere badituzte. Hizkuntzaren aldaketak sare sozialetan idatzizko lekukotza izaten du, beste inon baino bizkorrago. Eta lekukotza horiek erauz ditzakegu, era masiboan. Termino berrien agerpena, edo gazte hizkera bezalako erregistroen azterketa egiteko informazio iturri egokia izan daiteke, beraz (Nguyen, Gravel, Trieschnigg & Meder, 2013).

Iturri honetatik eratorri dezakegun informazioa baliatzeko adibide gisa, Twitter corpusean agertzen diren lemen azterketa egin dugu. Corpusaren lematizazioa beste corpusekin egin bezala Eustagger tresnarekin burutu dugu. Goian aipatutako aurreprozesatzerik ez dugu egin kasu honetan, gure helburua beste corpusekin konparatzea izan delako. Hortaz interes berezia genuen Sare sozialetan ematen diren fenomeno horiek azaleratzen. 2 eranskinean, hiru corpusetan agerpena duten lema-ikurren lagina eskaintzen dugu. Lema-ikurrek hiru corpusetan dituzten maiztasun erlatiboak (rfreq, ikus Lindemann & San Vicente (2015)) konparatzen ditugu, terminologia-erauzketan erabiltzen den weirdness ratio neurria erabiliz (Ahmad, Davies, Fulford & Rogers, 1994; ik. Schäfer, Rösiger, Heid & Dorna, 2015). Weirdness baloreak adierazten du ikur baten maiztasuna corpus batean, erreferentzia-corpus batean duenarekin alderaturik. Taulan zerrendatutako lema-ikurrek corpus zehatz batean askoz ere agerpen gehiago dute dagokion erreferentzia-corpusean baino, eta, hortaz, corpus zehatz horren osaeraren ikuspegi berezia eskaintzen dute. Ez dira corpus batean maizen agertzen direnak, erreferentzia-corpuseko maiztasunetik urrunenak gelditzen direnak baizik. Zerrendako lehenengo posizioei erreparatuta, Internetetik erauzitako edukiak biltzen dituzten Elh200 corpusak eta Twitter corpusak antzekotasuna erakusten dute, eskuz hautaturiko ereduzko testuez osatutako ETC corpusaren aldean. Bestalde, Twitter corpusak, Elh200 erreferentzia harturik, bere berezitasunak erakusten ditu, hemengo lehenengo tokietan ere erakuskari esanguratsuak agertzen baitira. Hemen aurkezten duguna oso oinarrizkoa bada ere, maiztasunak erkatzeko ariketa hau ezingo genuke burutu iturri ditugun baliabideak elkartu ahal izan ezean.

6. Ondorioak

Hiztegi historikoak eta bestelako baliabide lexikalak saretzea izan dugu hizpide artikulu honetan. Lan-urratsak ikusi ditugu, karaktereen ezagutza eta testu-edukien markaketa semantikotik datu-multzo lexikal egituratuez osatutako bilduma sortzeraino. Bestetik, horrelako bilduma sare semantikoekin erlazionatzeko baldintzak ere aztertu ditugu, horien artean, datu-ereduen estandarrei erreparatzeko premia.

Euskarazko hiztegi historikoei dagokienez, hastapen batzuk baino ezin izan ditugu gauzatu egun arte. Lemen aldaera historikoak egungo estandarrerantz moldatzea beharrezkoa dela ikusi dugu lema-ikurretan oinarritutako baliabide-sareak eraikitzeko. Beste puntu batean ere agertu zaigu premia bera: Sare sozialetan erabiltzen diren formak ere estandar diren edo liratekeen formekin lotzea. Lan-urrats horrek, aurrean ditugun aldaerak historikoak ala berri-berriak izanda, irizpide filologikoak eskatzen ditu. Euskarazko hiztegigintza historikoan lematizazioaz sortutako ereduak aztertu eta erabilgarri bihurtzea, corpus hizkuntzalaritzako tresnek aplika ditzaten, gerorako eginkizun garrantzitsua izan daitekeela aipatu nahi dugu, eta OEH-n lema historikoen kodifikatze lexikografikoaz egindako eskuzko lan erraldoia banan-banan erauztea horretarako bide egokia izan litekeela.

Datozen urteetan ildo horretan lanean jarraitzeko beharra eta aukera ikusten dugu. Elhuyar Fundazioa hasia da dagoeneko bide horretan. Batetik, fundazioak berak garatutako baliabide lexikalen inguruko loturak eskaintzen dizkie erabiltzaileei, corpusetatik eratorritako adibideak, edo hiztegi terminologikoetan egon daitezkeen sarreretara loturak. Horiek «barne loturak» direla esan dezakegu, erakunde berak sortutako baliabideen artekoak. Baina, Euskal Wikilarien Elkartearekin elkarlanean, Elhuyar hiztegietako datu lexikalak Wikidata ekimenean txertatzen hasiak dira. Euskarazko bestelako baliabideekin zein beste hizkuntzetako lexikalizazioekin loturak sortzeko pauso garrantzitsu bat eman da honezkero, aurrera begira, euskarazko hiztegien erabiltzaileari ere onurak ekarriko dizkiona, edozein bilaketa egitean, euskaraz ez ezik beste hainbat hizkuntzatako informazioa ere eskura izango duelako.

H2020 *Elexis* ikerketa-azpiegitura lagungarria izango zaigu, lan-fluxuak, datu-ereduak eta teknologiak Europako hizkuntza-komunitateetan zehar eztabaidatu, garatu eta zabaltzeko asmoa duena. Europako beste hainbat erakundek bezala, EHUk ere *Elexis* proiektuan hartzen du parte observer gisa, konsortzioari euskarazko datu lexikalak eskuragarri jarriz, azaldutako lan-fluxuaren urrats bakoitzean esperimentuak egin daitezen, eta esperimentu haien emaitzak ebaluatuz, tresna automatikoen garatzaileek ondorioak atera ditzaten. Trukean, garatutako teknologien early adopters izango gara. Tresnen garatzaileen eta datuen emaileen, informatikoen eta filologoen artean dugun nazioarteko lankidetza hori Humanitate Digitala arloa gorpuzten duten haietarikoa da, mundu biek irabazteko moduko lankidetza, zalantzarik gabe.[28]

[28] David Lindemannek Eusko Jaurlaritzaren laguntza eskertzen du (IT1169-19).

Bibliografia

Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Ezeiza, N., & Urizar, R. (1996). EUSLEM: A lemmatiser/tagger for Basque. Proceedings of EURALEX 1996, 17-26. Sarean: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.44.9004&rep=rep1&type=pdf

Ahmad, K., Davies, A., Fulford, H. & Rogers, M. (1994). What is a term?: The semi-automatic extraction of terms from text. In M. Snell-Hornby, F. Pöchhacker & K. Kaindl (Arg.), Benjamins Translation Library (Libk. 2, or. 267). https://doi.org/10.1075/btl.2.33ahm

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernandez, G. & Lersundi, M. (2001). EDBL: A General Lexical Basis for the Automatic Processing of Basque. *Proceedings of the IRCS Workshop on linguistic databases*, Philadelphia. Sarean: http://artxiker.ccsd.cnrs.fr/docs/00/08/11/54/PDF/2001-IRCS.pdf

Alegria, I., Aranberri, N., Comas, P. R., Fresno, V., Gamallo, P., Padró, L., San Vicente, I., Turmo, J., Zubiaga, A. (2015). TweetNorm: A Benchmark for Lexical Normalization of Spanish Tweets. *Language Resources and Evaluation*, *49*(4), 883-905. Sarean: https://doi.org/10.1007/s10579-015-9315-6

Artola, X., Ezeiza, N., Gurrutxaga, A., Sagarna, A. & Urkia, M. (2017). Lexikoaren Behatokia: Leiho bat XXI. mendeko hedabideetako euskarari. *Senez: itzulpen aldizkaria*, (48), 16. Sarean: https://dialnet.unirioja.es/servlet/articulo?codigo=6206911

Azkarate, M., Kintana, X. & Mendiguren, X. (Arg.). (2006). Elhuyar hiztegia: Euskara-gaztelania, castellano-*vasco* (3. arg.). Usurbil Gipuzkoa: Elhuyar Edizioak. Sareko interfazea: http://hiztegiak.elhuyar.org

Azkarate, M., & Lindemann, D. (2018). Basque Lexicography and Purism. *International Journal of Lexicography*, *31*(2), 132-150. Sarean: https://doi.org/10.1093/ijl/ecy003

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. *Proceedings of the The 51st Annual Meeting of the Association for Computational Linguistics*, 1352-1362. Sarean: http://anthology.aclweb.org/P/P13/P13-1133.pdf

Camacho-Collados, J. & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, *63*, 743-788. Sarean: https://jair.org/index.php/jair/article/view/11259

Declerck, T., McCrae, J., Navigli, R., Zaytseva, K. & Wissik, T. (2018). ELEXIS - European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data. In I. Kernerman & S. Krek (Arg.), *Proceedings of the LREC 2018 Workshop «Globalex 2018 – Lexicography & WordNets»* (or. 17-22). Sarean: https://globalex.link/globalex2018/proceedings/

Euskaltzaindia. (2010). Hiztegi batua (3.). Sarean: http://www.euskaltzaindia.net/hiztegibatua/

Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Arg.), *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*. Sarean: https://elex.link/elex2017/proceedings-download/

Larramendi, M. (1745). Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa. Sarean: http://www.kmliburutegia.eus/Record/203133 eta http://www.kmliburutegia.eus/Record/26577

Leturia, I. (2014). *The Web as a Corpus of Basque* (PhD Thesis). UPV-EHU Lengoaia eta Sistema Informatikoak Saila, Donostia. Sarean: https://www.elhuyar.eus/media/content/files/ILeturia-The%20Web%20as%20a%20Corpus%20of%20Basque.pdf

Lindemann, D. & Kliche, F. (2017). Bilingual Dictionary Drafting: Bootstrapping WordNet and BabelNet. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Arg.), Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017 (or. 23-42). Sarean: https://elex.link/elex2017/wp-content/uploads/2017/09/paper02.pdf

Lindemann, D. & San Vicente, I. (2015). Euskarazko maiztasun lemategia gaurko teknologien ikuspuntutik. In B. Fernández Fernández & P. Salaburu Etxeberria (Arg.), *Ibon Sarasola, gorazarre. Homenatge, homenaje* (or. 441-456). Sarean: http://www.ehu.eus/ehg/sarasola/liburua/SarasolaGorazarre36.pdf

Lindemann, D. & San Vicente, I. (2016). Bilingual Dictionary Drafting: Connecting Basque word senses to multilingual equivalents. In T. Margalitadze & G. Meladze (Arg.), Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity (or. 898-905). Sarean: http://euralex.org/publications/bilingual-dictionary-drafting-connecting-basque-word-senses-to-multilingual-equivalents/

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S. & Kedzia, P. (2016). plWordNet 3.0—A Comprehensive Lexical-Semantic Resource. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, 2259-2268. Sarean: https://www.aclweb.org/anthology/C/C16/C16-1213.pdf

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, *3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Mitxelena, K. & Sarasola, I. (1988). Diccionario general vasco—Orotariko euskal hiztegia. Euskaltzaindia; Editorial Desclée de Brouwer. Sarean: https://www.euskaltzaindia.eus/hizkuntza-baliabideak/online-hiztegiak

Navigli, R. & Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 216-225. Sarean: http://dl.acm.org/citation.cfm?id=1858681.1858704

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). «How old do you think I am?» A study of language and age in Twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. -(E)n aurkeztua Cambridge MA. Sarean: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPaper/5984

Pedersen, B. S., Nimb, S., Olsen, S. & Sørensen, N. H. (2018). Combining Dictionaries, Wordnets and other Lexical Resources – Advantages and Challenges. In I. Kernerman & S. Krek (Arg.), *Proceedings of the LREC 2018 Workshop «Globalex 2018 – Lexicography & WordNets»* (or. 101-104). Sarean: https://globalex.link/globalex2018/proceedings/

Pociello, E. (2007). *Euskararen ezagutza-base lexikala: Euskal WordNet* (PhD Thesis). UPV/EHU Euskal Filologia Saila, Donostia. Sarean: http://ixa.si.ehu.es/node/4117

Raganato, A., Camacho-Collados, J. & Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 99-110. Sarean: https://www.aclweb.org/anthology/E17-1010/

Sarasola, I. (1996). Euskal Hiztegia. Donostia: Kutxa Gizarte-eta Kultur Fundazioa.

Sarasola, I. (2008). Egungo Euskararen Hiztegia (EEH). Berreskuratua 2019(e)ko azaroakaren 12a, -(e)tik UPV/EHU website: https://www.ehu.eus/eeh/

Sarasola, I., Landa, J. & Salaburu, P. (2013). Egungo Testuen Corpusa. Sarean: http://www.ehu.es/etc/

Saurí, R., Mahon, L., Russo, I. & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier. In M. Eskevich, G. de Melo, C. Fäth, J. P. McCrae, P. Buitelaar, C. Chiarcos, M. Dojchinovski (Arg.), *2nd Conference on Language, Data and Knowledge (LDK 2019)* (or. 20:1-20:16). https://doi.org/10.4230/OASIcs.LDK.2019.20

Schäfer, J., Rösiger, I., Heid, U. & Dorna, M. (2015). Evaluating noise reduction strategies for terminology extraction. In T. Poibeau & P. Faber (Arg.), *Proceedings of the 11th Int. Conference on Terminology and Artificial Intelligence (TIA'15)* (or. 10). Granada: Universidad de Granada. Sarean: http://ceur-ws.org/Vol-1495/paper_7.pdf

Smith, M. N. (2004). Electronic Scholarly Editing. In S. Schreibman, R. Siemens & J. Unsworth (Arg.), *A Companion to Digital Humanities*. Sarean: http://www.digitalhumanities.org/companion/

1. eranskina

Sarasolaren (Larramendiren) «*1745» lemen maiztasun-zerrenda, ETC eta Elh200 corpusetik erauzitako maiztasun-zerrendan dituen tokiak adierazita (40 lehenengoak).

Posizioa	Lema	ETC rank	Lema	Elh200 rank
1	talde	39	talde	42
2	erabaki	100	aurkeztu	93
3	egoera	121	ondoren	95
4	azaldu	128	elkarte	108
5	maila	136	erabaki	120
6	ondoren	143	maila	130
7	biztanle	174	egoera	134
8	arazo	183	azaldu	173
9	aurkeztu	207	era	208
10	elkarte	233	arazo	215
11	asmo	241	atal	268
12	era	281	irail	271
13	arrazoi	301	asmo	311
14	saiatu	355	honako	319
15	zuzendari	373	zehar	348
16	sarrera	381	hona	359
17	bukatu	394	sarrera	364
18	ordaindu	423	bultzatu	367
19	bultzatu	455	albiste	379
20	ondorengo	459	ordaindu	395
21	adina	510	martxa	411
22	doan	543	arrazoi	412
23	zehar	545	zuzendari	416
24	atal	549	bukatu	444
25	giro	550	lasai	512
26	egitura	559	saiatu	526
27	baztertu	560	hitzaldi	538
28	lasai	565	hondakin	559
29	antzeko	569	ale	581
30	desagertu	615	egitura	587
31	irail	630	bideratu	617
32	iparralde	641	giro	620
33	garaipen	651	arautu	640
34	zalantza	664	gehitu	649
35	honako	693	ipuin	652
36	aski	700	zalantza	675
37	benetan	716	bizitu	685
38	hainbeste	719	egiaztatu	710
39	sasoi	721	neu	712
40	proba	747	antzeko	721

2. eranskina

Twitter corpusean gutxienez hogei aldiz nahiz ETC eta Elhuyar corpus bietan agertzen diren lema-ikurrak, weirdness-ratio neurriaren arabera ordenaturik, erreferentzia-corpus ezberdinak oinarritzat hartuta (100 lehenengoak). Hirugarren eta laugarren zutabeetan, Twitter corpuseko maiztasun erlatiboak ETC eta Elh200 corpusetako maiztasunekin, eta bostgarren zutabean, berriz, bien batezbestekoarekin duen proportzioaren araberako hurrenkera dugu, hau da hirugarren zutaberako, Twitteren soilik internet osoan baino maiztasun askozaz ere handiagoaz erabiltzen direnen erakuskaria.

Posizioa	Elh200_rfreq / ETC_rfreq	Twitter_rfreq / ETC_rfreq	Twitter_rfreq / Elh200_rfreq	Twitter_rfreq / ETC_Elh200_rfreq
1	batera	batera	ezker	etxeondo
2	arteko	noranzko	mikroskopiko	zapra
3	aurrez	arteko	aspirazio	patatx
4	bertatik	zuzeneko	astintzaile	antifaxismo
5	arriskutsu	arriskutsu	berritsukeria	manugaitz
6	baserritar	bertatik	botagura	torraka
7	zuzeneko	baserritar	burugorri	egibide
8	arrakastatsu	aurrez	indargetu	txantxibiri
9	bigarren	arrakastatsu	saharaui	infrasoinu
10	inon	bigarren	amatista	txerren
11	zehazki	inon	tematu	elikagune
12	berrikusi	eguberri	interrogatorio	mozal
13	garapa	urrezko	gedar	salabardo
14	zelan	aspaldiko	librepentsalari	noranzko
15	jatorriz	etxeondo	patroi	mertzero
16	zeharka	zelan	tiratu	eztabai
17	aspaldiko	ospa	artalde	enkantatu
18	bada	orduko	tesla	prospero
19	urrezko	indartsu	frustazio	nabaro
20	bertako	berrikusi	tuneatu	spoiler
21	zai	zai	txeketegi	bermiotar
22	indartsu	fuerte	egozentriko	zerraldatu
23	hirugarren	txiolari	kontari	golatu
24	bakarka	zehazki	fiskalizatu	brutal
25	orduko	koloretsu	pito	azarri
26	eskuko	larriki	bikini	berjabetu
27	berreskuratu	hamargarren	hispanoamerikar	euskarafobo

Posizioa	Elh200_rfreq / ETC_rfreq	Twitter_rfreq / ETC_rfreq	Twitter_rfreq / Elh200_rfreq	Twitter_rfreq / ETC_Elh200_rfreq
28	berreraiki	klaru	pertso	behizain
29	berriki	berreskuratu	hirixka	zibereraso
30	nora	zabalik	tupla	proximo
31	gogoko	eskuko	bengaliar	armagabetu
32	ageriko	bizirik	graduatu	fronton
33	bizirik	nora	marigorringo	basomutil
34	ahozko	gogoko	normalitate	txiotxio
35	nahikoa	bertako	enbaxada	horrible
36	hamargarren	garapa	kultista	bonito
37	hare	hirugarren	arrastiri	alegrias
38	neu	ikusarazi	israeldar	sorki
39	klaru	zeharka	dibisa	jornada
40	gehiegizko	oriotar	domeinatu	txiolari
41	amen	ageriko	arilo	lagunkoitasun
42	gaineko	bada	zorun	playoff
43	onuradun	mago	zabalketa	erretura
44	interesdun	keatu	zakuto	restaurante
45	koloretsu	aitzina	konopial	bidelari
46	misteriotsu	basura	dolore	desakatu
47	zatika	hagoatu	irudikeria	maravilloso
48	delako	berreraiki	atralaka	salio
49	zentsu	gogotsu	hamargarren	estatalizazio
50	berregin	ikustarazi	espainiarzale	bypass
51	zabalik	hiruko	gexo	habuin
52	baliotsu	neu	zentralizatzaile	larretxe
53	berraztertu	biziarazi	armagabe	taup
54	eguberri	ustezko	momotxorro	inguralde
55	zeia	bakarka	indigaztainondo	migrante
56	dirudienez	barrura	txortalo	oholtzape
57	zegi	jatorriz	ezpatadantza	fototeka
58	guztizko	irundar	glass	movil
59	kartzeatu	adimentsu	tentaldi	zalagarda
60	ustezko	polideportibo	mahi	ziztor
61	legezko	nahikoa	ur	mineri

Posizioa	Elh200_rfreq / ETC_rfreq	Twitter_rfreq / ETC_rfreq	Twitter_rfreq / Elh200_rfreq	Twitter_rfreq / ETC_Elh200_rfreq
62	baketsu	baliotsu	trikatu	txinta
63	maket	berregin	ingurukatu	baluarte
64	saltoka	erdutu	izain	txamarreta
65	garrantzitsu	delako	tease	cadiztar
66	lapurtar	kartzeatu	arazle	oxala
67	mutxi	dirudienez	hargindegi	ugaitz
68	esplikazio	berriki	beroketa	traol
69	belasko	aberia	beno	ikasmin
70	ezagutarazi	belasko	hondeagailu	ahi
71	ikusarazi	gehiegizko	begigorri	kurkubi
72	gutxiegi	saltoka	tridente	mordaza
73	historikoki	lapurtar	sprinter	euskaitz
74	oriotar	haizegune	ondoezik	txotor
75	formalki	misteriotsu	hondarpe	kupe
76	sutsu	baldintzape	ikusketa	aurrestreinaldi
77	odoltsu	zeu	goizale	buruto
78	kementsu	baketsu	andetar	urgatzi
79	geroko	zentsu	jiratu	ihitza
80	moduzko	gutxiegi	txamarra	artefaktu
81	eregi	laguntxo	lorezale	twit
82	gehi	amen	relazio	hankazabal
83	aitzina	pentsiodun	apiril	turf
84	apurka	dozenaka	elkarreragile	haizeberritu
85	baiezko	maket	petraldu	animos
86	zeu	xitu	itsusikeria	gizaeskubide
87	zazpigarren	pago	mandatu	polideportibo
88	halakoxe	garrantzitsu	alergeno	moreno
89	berridatzi	ahozko	berantetsi	tanda
90	bertsu	zapra	sintetiko	puto
91	borondatezko	zazpigarren	errekurrente	galgorri
92	larriki	joko	zarpaildu	saratu
93	ereduzko	trabela	apurkor	argazkigile
94	berrantolatu	gaineko	ezkutaleku	semifinal
95	giza	kostata	klasifikazio	estreno

Posizioa	Elh200_rfreq / ETC_rfreq	Twitter_rfreq / ETC_rfreq	Twitter_rfreq / Elh200_rfreq	Twitter_rfreq / ETC_Elh200_rfreq
96	pasonibel	pina	barrendero	ahaldungai
97	berregituratu	giza	zarabanda	rikar
98	orokorki	sutsu	itsasgora	gipuzkoano
99	adimentsu	pozoitsu	kontrapuntu	jenial
100	latindar	baiezko	esperoan	sorteo