EUSKARAREN NORMALIZAZIOA
ETA
LINGUISTIKA KONPUTAZIONALA

Itziar ADURIZ
Izaskun ALDEZABAL
Joxe Mari ARRIOLA
Xabier ARTOLA
Miriam URKIA

EUSKARAREN NORMALIZAZIOA ETA LINGUISTIKA KONPUTAZIONALA[1]

I. Aduriz, I. Aldezabal, J.M. Arriola, X. Artola, M. Urkia

INFORMATIKA ETA HIZKUNTZA

Txosten honen helburu orokorra, euskararen normalizazioari eragiteko, sistematizazio eta formalizaziorako tresna lagungarrien berri ematea da, linguistika konputazionalaren ildotik.

Linguistika konputazionalak edo, nahi bada, lengoaia naturalaren prozesamenduak hizkuntza hartzen du gai eta ordenadorea lanabes. Konputagailuz hizkuntza ulertu eta sortzea, edo hizkuntza batetik bestera itzulpenak egitea, edota, esate baterako, gizaki-makina elkarrizketetako hariei igartzea, horrelakoxeak dira arlo honetan dabiltzan ikertzaileen asmo eta kezkak.

Horiek guztiak taxuz landuko badira derrigorrezkoa da hizkuntzaren azterketa sistematizatzea eta normalizazio-bidean aurrera jotzea. Guri behintzat urteotako eskarmentuak hori erakutsi digu. Uste dugu, gainera, sakondu behar horrek bi alde dituela: batetik, ordenadorea lagun, orain dela gutxi ezinezkoak ziruditen azterbideak egingarriak dira eta, bestetik, hizkuntz trataera automatizatzeko ahaleginak berarekin dakar hizkuntzaren deskribapena formalizatu eta doitu beharra. Lehenengo aldeak hizkuntzalariari lagundu egingo dio, bigarrenak, akuiluak bezala, behartu egingo du.

Aspalditxotik hasi ziren informatikari eta hizkuntzalariak burutazio hauei forma ematen. Hasieran, 60. hamarkadan, ilusioz beteta, burua baino gogoa arinago, goi-mailako itzulpen automatikoa laster lortu zitekeelakoan. Gero, etsipenak jota, ez gogo, ez ilusio. Azken urteotan, berriz, aplikazio erabilgarriak egin nahi izan dira, nahiz eta bistakoa den hizkuntza -lengoaia naturala- moldagaitza dena. Baina, hala ere, badira gaur egungo merkatuan aplikazio interesgarriak, batzuk oso eskura daudenak: hizkuntz egiaztatzaile/zuzentzaileak, terminologi datu-baseak, hiztegi elektronikoak, etab.

INFORMATIKA ETA EUSKARA

Gure taldearen erronka joera berri hauetara euskara egokitzea da. Helburu horrek ia hutsetik hasi beharra ekartzen du batetik eta, euskararen berezitasunak direla eta, zenbait problema desberdini aurre egin beharra bestetik. Gauzak horrela, azpiegitura sendo bat eratzen

[1] Txosten honek UZEI eta Donostiako Informatika Fakultateko (EHU) LNP ikertaldea elkarlanean garatzen ari diren proiektuen aurkezpena izan nahi du.

saiatu gara. Emaitza edo produktu ikusgarrietara jotzea baino nahiago izan dugu ezari-ezarian hasi eta norabide egokian pausoak ematea.

Aipatu azpiegiturak bi zutabe ditu: oinarrizko datu-base lexikala eta testuz nahiz hiztegiz osatutako korpusa. Biltegi horiek laguntza paregabea ematen dute euskararen izaera aztertzeko, izan ere, halako azterketarik gabe prozesamendu automatikoa herren geratzen baita. Euskarria osatzeko analizatzaile/sortzaile morfologiko eta sintaktikoak beharrezkoak dira. Morfologia landu zen aurrena, sintaxiarekin dihardugu egun eta semantika dugu gure asmoen artean.

EUSKARARAKO APLIKAZIOAK

Ingurune honen barruan garatutako eta garatzen ari garen proiektuen aurkezpentxoa da ondoko lerroetan dakarguna, axalekoa bada ere.

0. HAIN: Hizkuntz Aplikazioetarako INgurunea

Proiektu honen helburua hizkuntz aplikazioetarako ingurune informatiko bat lortzea da, alegia, xede desberdinetarako hizkuntz tresnaz osaturiko multzo bat era integratu batez biltzea. Ingurune honek irekia izan nahi du eta edozein erabiltzailerentzat eskura egon, modu atseginez kontsultatu ahal izateko.

Ondoren aurkeztuko ditugun moduluek osatzen dute aipatu ingurunea:

1. EDBL (Euskararako Datu-Base Lexikala)

EDBL, lexikoari buruzko informazio-biltegi erraldoia den neurrian, helburu askotarako erabilgarria izan daiteke [Agirre eta beste 94a]. Gure interesa, oraingoz, euskararen tratamendu automatikoari dagokiona da. Beraz, EDBL euskararen prozesamendu automatikoan zeregin desberdin askotarako oinarri lexikala da, hau da, arloko lan desberdinetan beharrezko diren lexikoien oinarri eta iturri.

50.000 sarreratik gora ditu datu-baseak, bakoitza bere kategoria eta informazio morfologikoarekin. Funtzioari buruzko hainbat informazio ere gehitu da, sintaxia eta semantikarako bidea irekiz. Hiztegi moderno eta zabala da, euskara batua oinarrian duena eta Euskaltzaindiaren erabaki/gomendioekin batera eguneratzen dena.

EDBLk honako ezaugarriak ditu: **helburu askotarako tresna** (analisi morfologikorako bezala beste mailetako analisiak egiteko oinarri egokia); **neutrala** (ez du formalismo jakin bati estuki lotuta egon behar); **malgua eta irekia** (edozein momentutan helburu berrietarako egokitzen erraza, informazio-mota berriak aise onartzen dituena) eta **erabilerraza** (erabiltzaile arruntari zein berezituari behar bezalako laguntzak eta erraztasunak eskaintzen dizkiona).

Bertan informazio guztia taula desberdinetan banatua dago ondoko adibidean errepresentatuta ikus dezakegun bezala:

```
TN_HS    (Sarrera, Homografo-identifikadorea, Kategoria, Forma flexionatua,
  Hitz anitzeko sarrera, Eratorria, Hitz-elkarketa, Iturburua, Iturburuko
  forma)

      (idazmakina, 1, IZE, -, -, -, +, K, idazmakina)


BIM_FORM  (Sarrera, Homografo-identifikadorea, BIM-Forma, Jarraitze-
  klasea, Lexikoia)

      (idazmakina, 1, idazmakinA, I, izenak)


IMS_IZE (Sarrera, Homografo-identifikadorea, Azpikategoria, Biziduna,
  Zenbakarria, Neurgarria, Plurala)

      (idazmakina, 1, ARR, -, +, -, -)


IMS_ELK (Sarrera, Homografo-identifikadorea, Mugakizuna, Mugakizureren
  Homografo-identifikadorea, Mugatzailea, Mugatzailearen Homografo-
  identifikadorea, Elkarketa-mota)

      (idazmakina, 1, makina, 1, idatzi, 1, ADI+I)
```

2. Analizatzaile Morfologikoa

Euskararen morfologia konplexua kontuan izanda (eranskaria) eta gaurko egoera ikusirik (batu berria, formalizatu gabea ikuspegi konputazionaletik behintzat), beste hizkuntzetan egindakoa aztertu eta Koskenniemi-ren [Koskenniemi 83] bi mailatako morfologiara jo dugu euskal morfologiaren tratamendu automatikoa gauzatu ahal izateko [Agirre eta beste 89], [Agirre eta beste 91].

Formalismo honen ezaugarri nagusiak bat datoz gure beharrekin: a) eredu orokorra da, edozein hizkuntzari aplika dakiokeena; b) baliagarria da hitzen analisi morfologikorako zein hitz sorkuntzarako; c) ezagutza linguistikoa eta algoritmoa bereiz daude, eta ondorioz, programa berak edozein hizkuntzatarako balio dezake; d) analizatu edo sortuko den hitzaren azaleko maila eta sakonekoa argi eta garbi bereizten ditu; ondorioz, ez dago aldaketa morfofonologikoengatik sortutako morfema baten forma desberdinak gorde beharrik; e) fonologia sortzaileko berridazketa-erregelen ordez, erregela paraleloak erabiltzen ditu, sistema kontzeptual zein konputazionalki errazago bihurtuz.

Oinarrizko osagaiak **sistema lexikoa** (lexikoa bera bere jarraitze-klaseekin, hau da, lemaren atzean joan daitezkeen morfema-multzoekin) eta **erregelak** dira.

Azkenean lortzen den analisiaren emaitza horrelakoa da:

```
I Z E N A : amak
((forma "amak")
  ((anal 1)
   ((lema "amA")((KAT IZE)))
   ((morf "ak")((KAT DEK)(KAS NOM)(NUM P)(MUG M))))
  ((anal 2)
   ((lema "amA")((KAT IZE)))
   ((morf "ak")((KAT DEK)(KAS ERG)(NUM S)(MUG M))))
  ((anal 3)
   ((lema "amA")((KAT IZE)))
   ((morf "Ek")((KAT DEK)(KAS ERG)(MUG MG))))
)
```

3. XUXEN: Egiaztatzaile/Zuzentzaile Ortografiko Automatikoa

Goian aipatutako bi proiektuek albo-produktuak sortzeko aukera eman digute. Denetan behinena XUXEN egiaztatzaile/zuzentzaile ortografikoa izan da [Agirre eta beste 92].

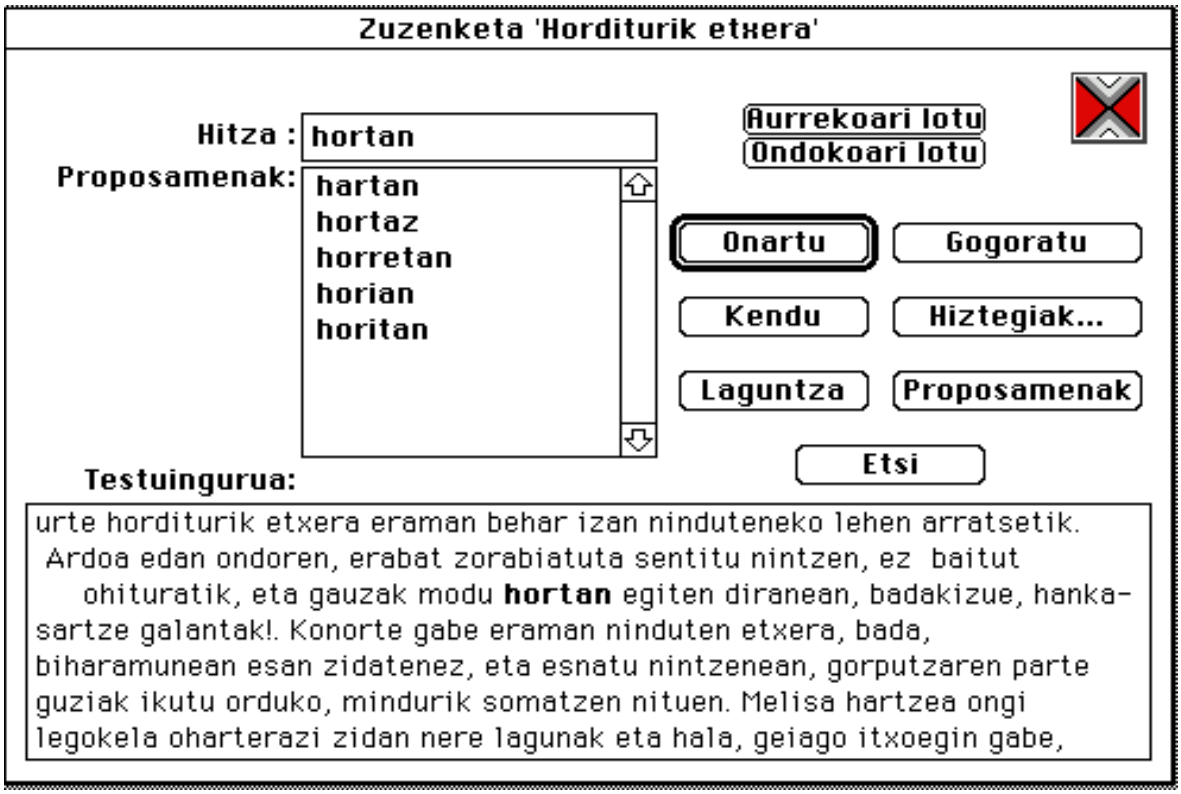
Analisi morfologikoan oinarrituta dagoenez, hitz bat aintzakotzat hartzen da analisi posiblerik badu, eta bestela errore dela kontsideratuko du eta zuzentzeko prozesuan sartu. Oraingoz XUXENek ez du errore sintaktiko edo semantikorik detektatzen, hitz-forma hartzen baitu aztergai. Azaltzen diren erroreak ezjakintasunak eragindakoak izan daitezke (erregela eta azpilexiko berezien bidez konpontzen direnak[2]) edo tipografikoak (trigramen konbinazioen bidez konpontzen direnak).

EDBLko sarrerek osatzen dute hiztegia, nahiz erabiltzailearen hiztegi pertsonala sortzea ere posible den zuzendu nahi den testua oso espezializatua denean (medikuntza, kimika...).

Testu baten zuzenketa-prozesuan XUXENek hartzen duen itxura honako hau da, zuzen ez dagoen hitzaren ordezko aukerak eta guzti:

---

[2] Hizkuntzaren ezagutzari erreparatuz (ezagutza-eza normalizazioaren eraginarengatik, forma dialektalak erabiltzeagatik,...) zenbait errore sarri gertatzen dira. Horiek aztertu eta sistematizazio-saio bat egin dugu bi mailatako azpisail paraleloa sortuz, gure sisteman egoki integratu ahal izateko eta sistemak berak horretarako aukera ematen zigulako.



4. Analisi sintaktikoaren bideetan

Bada hainbat lan aipagarri euskal gramatikaren inguruan (Euskaltzaindiaren *EGLU* [Euskaltzaindia 85], Patxi Goenagaren [Goenaga 80] *Gramatika Bideetan*, etab.), baina maila deskriptiboan mugitzen dira eta analisi automatikorako erregela finkoak dira beharrezkoak. Maila morfologikoan ez bezala, sintaxian horren falta nabaria sumatu dugu.

Sistematizazio konputazionalerako *Murriztapen-Gramatikak* [Karlsson eta beste 92] izeneko formalismoan oinarritu gara, euskararen kasurako egokiena delakoan. Testuingurua kontuan hartzeaz gain, informazio morfologikoan oinarritzen da analisirako eta desanbiguaziorako. Testu errealak helburu dituenez, korpusetatik abiatuta eraikitzen dira erregelak.

Lehenengo urratsak eman dira, oinarrizkoenak nahi bada, baina poliki goaz ezinbestean. Oraingoz desanbiguazioa (morfologia mailakoa, kategoriala, paradigmatik kanpokoa eta sintaktikoa) eta funtzioen asignazio sintaktikoak landu ditugu, hurrengo azterketarako oinarri izango direnak.

Hona hemen formalismo honekin lortuko dugun analisi sintaktikoaren itxura:

```
itsasoak ez du esperantzarik.

((forma "itsasoak")
  ((anal 1)
     ((lema "itsaso")((KAT IZE)))
     ((morf "ak")((KAT DEK)(KAS NOM)(NUM P)(MUG M))))
  ((anal 2)                                                  @SUJ (@NORK)
     ((lema "itsaso")(KAT IZE)))
     ((morf "ak")((KAT DEK)(KAS ERG)(NUM S)(MUG M))))
)

((forma "ez")
  ((anal 1)                                                  @EZ_PART
     ((lema "ez")((KAT ADB))))
  ((anal 2)
     ((lema "ez")((KAT IZE))))
  ((anal 3)
     ((lema "ez")((KAT IZE)))
     ((morf "0")((KAT DEK)(KAS NOM)(MUG MG))))
)

((forma "du")
  ((anal 1)                                                  @+JADNAG
     ((lema "duE")((KAT ADL)(MDN A1)(NOR 3)(NRK 3)(ERR *edun))))
)

((forma "esperantzarik")
  ((anal 1)                                                  @OBJ
     ((lema "esperantzA")((KAT IZE)))
     ((morf "Rik")((KAT DEK)(KAS PAR)(MUG MG))))
)

((forma ".")
)
```

Sistema honen aplikazio nagusietako bat erroreak detektatu eta zuzentzeko oinarriak finkatzea izango da. Bestalde, baliagarria izango da testuen azterketarako, hizkuntzen irakaskuntzarako eta estilo-zuzentzaile gisa (idazketarako laguntza-sistemak), euskararen tratamendu automatikoa beharrezko duten beste aplikazio aurreratu askotarako bezalaxe.

5. Informazio lexiko-semantikoaren erabilera

Apenas landu den gaia dugu honako hau. Azpikategorizazioa osatu nahirik, hautapen-murriztapenak ezartzeari ekin diogu.

Batetik, hitzaren argumentuek bete beharreko ezaugarri semantikoak adierazten dituzten hautapen-murriztapenak ezarri ditugu. Hartara, *apurtu* hitzaren azpikategorizazioa kontuan hartuta, zuzentzailea gai izango da "lehio bat apurtu dut" esaldiko *lehio* hitz okerrari dagokion zuzena *lehio: lehia, leiho, lesio* proposamenen artetik bereizteko.

Bestetik, formalismoaren arabera, hierarkia semantiko baten barruan hitzen adieren arteko distantzia semantikoa neurtuko da. Beraz, *itsasuntzi* eta *kapitain* formen arteko distantzia "hurbila" izango da eta *itsasuntzi* eta *negar* formen artekoa, berriz, "urruna" [Agirre eta beste 94b].

6. Zuzentzaile aurreratua

3. atalean aipatu bezala, zuzentzaileak ez ditu errore sintaktiko edo semantikoak tratatzen, hitzaren analisira mugatzen da. Baina gure erronka testuingurua kontuan hartuko duen analizatzaile eta zuzentzailea bideratzea da [Gojenola 94]. Horretarako ezinbestekoak dira analizatzaile sintaktikoa (4. atala) eta semantikoa (5. atala).

Egun XUXENek (3. atala) ontzat ematen ditu *Joseba liburua ekarri du* edo *Gezurra eman dit* bezalakoak, testu-hitzak banaka gramatikalak baitira. Baina ergatibitatea eta azpikategorizazioa ez dira lehen maila honetan landu. Goiko puntuak landu ahala, zuzentzaile aurreratua posible izango da [Abaitua eta beste 92].

7. EUSLEM: EUSkararako LEMatizatzaile automatikoa

Proiektu honen helburua lematizatzaile/etiketatzaile sendoa eta erabilpen orokorrekoa (lexikografia, datu-base dokumentalak, sintaxia...) sortzea da, hitz-formetatik lemak lortu eta horrekin batera kategoria esleituko duen tresna [Aduriz eta beste 94]. Horretarako oinarri bezala erabili dira EDBL eta analizatzaile morfologikoa, bertan errore tipikoen tratamendua erantsi delarik.

Aurreko proiektuetan bezala, hemen ere ez dira formalizazio-arazoak falta izan, etiketa-sistemaren aldetik eta, orokorrean, oinarri linguistikoen aldetik. Teorizazio- eta sistematizazio-lan handia egin da gai honen inguruan.

Behar guztiei erantzuteko hitz anitzeko terminoek eskatzen duten tratamendu berezia aztertu da, lexiko gabeko lematizazioa prestatu eta estatistiketan oinarritutako desanbiguazio automatikoa ere landu da. Bestalde, zenbakiak, laburdurak, puntuazio-ikurrak, etab. aurreprozesatzen dituen tresna bat garatu da.

TESTU-HITZA	LEMA	ALDAERA LEMATIZATUA	KATEGORIA
testuko	testu	testu	IZE
baltzaren	beltz	baltz	ADJ
zuaitzeko	zuhaitz	zuaitz	IZE

8. Beste proiektu batzuk

Orain arte aipatutakoez gain badira lengoaia naturalaren prozesamenduaren inguruan beste proiektu edo asmo interesgarri batzuk, hauek ere euskararen tratamendu automatikoan lagungarri izan daitezkeenak:

8.1. Lexikografia konputazionala (Agirre, E., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Sarasola, K., Soroa, A.): arlo honen barruan eraikitako hiztegi-sistema urgazle adimendunak (HIZTSUA [Artola 93] eta ANHITZ [Agirre eta beste 94c]) dira orain arte egindako lanak, hiztegiko ezagutza era inteligente batean ustiatu ahal izateko errepresentatuta daudenak.

HIZTSUA hiztegi elebakarra den bezala, ANHITZ itzulpenerako laguntza-sistema da, hots, hiztegi elebidunetan pentsatua dago.

8.2. Elkarrizketen azterketa [Zubizarreta 93]**:** pragmatika, diskurtso-azterketaren alorrean kokatua, gramatikak esaldiaren mailatik elkarrizketaren mailara eramatea du helburu, hau da, elkarrizketaren gramatika funtzional bat garatu da. Domeinutzat argazkigintza hartu da.

8.3. Bigarren hizkuntzaren ikaskuntza-prozesuaren azterketarako sistema laguntzaile adimenduna (Díaz de Ilarraza, A., Maritxalar, M.): linguistika, psikolinguistika eta adimen artifiziala biltzen dituen sistema honekin ikaskuntza-prozesuan gertatzen diren fenomeno konkretu batzuen azterketa egiten da (tarte-hizkuntza, fosilizazioa eta ikaslearen zenbait ikaste-estrategiaren eragina aurreko bietan). Horretarako euskaltegietako ikasleen testuak erabiltzen dira eta elkarlana suspertu.

8.4. Korpusaren estandarizazioa eta aberastea (Alegria, I., Ezeiza, N.): Europa mailan erabiltzen ari diren estandarrak eredutzat hartuta, biltegituta dauzkagunak haietara egokitu eta testu gehiago gorde nahi dugu.

8.5. AHOZKA (Hernaez, I., Etxeberria, P., Txillardegi): ahotsa sintetizatzeko garatzen ari den proiektua. Makinen eta gizakien arteko harremanetan makinak gizakiari eman beharreko mezuak idatzizko hizkuntzaz eman ordez ahozkoaz ematea da iharduera honen oinarrizko helburua.

ONDORIOAK

Aurreko puntuetan euskararen tratamendu automatikoa bideratu ahal izateko beharrezkoak diren tresna eta moduluen aurkezpentxoa egin da eta zein ingurune eraiki den aipatu. Baina ez da automatizazioa helburu bakarra; izan ere, horren aurretik hizkuntzak normalizatua behar du eta hutsune hori sumatu dugu zenbaitetan, gehienetan deskribapen hutsa aurkitu dugulako eta sistematizaziorako ez da nahikoa. Horregatik, gure saioa hankamotz gera daiteke.

Gure aurkezpena egiteko aukera eman zaigunez, ez genuke aipatu gabe utzi nahi diziplina arteko elkarlanaren beharra, gramatika finkatu ahala arlo guztietara aplikatu ahal izango dugulako eta, azkenean, euskara bera ere automatikoki tratatzea amets baino errealitate izango delako.

Bestalde, tresna hauek, euskara lantresna duen edonorentzat baliagarri ere badira, gure altxorra biltegiratua edukitzea eta zuzenean ustiatzea ezinbestekoa baita luza gabe.

ERREFERENTZIAK

[Abaitua eta beste 92] Abaitua, J., Aduriz, I., Agirre, E., Alegria, I., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M., Sarasola, K., Urkia, M., Zubizarreta, J.R.: *Estudio comparativo de diferentes formalismos sintácticos para su aplicación al euskara* (barne-txostena UPV/EHU/LSI/TR-193), 1992

[Aduriz eta beste 94] Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Urkia, M.: "EUSLEM: Un lematizador/etiquetador de textos en euskara", in *SEPLN*, Córdoba, 1994.

[Agirre eta beste 89] Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K., Urkia, M.: "Aplicación de la morfología de dos niveles al euskara", in *SEPLN (8)*, Barcelona, 87-102, 1989.

[Agirre eta beste 91] Agirre, A., Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Goenaga, P., Maritxalar, M., Sarasola, K., Urkia, M.: "Bi mailatako morfologiaren euskararako egokitzapena", in *Elhuyar*, 1991, 7-14.

[Agirre eta beste 92] Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K., Urkia, M.: "XUXEN: A spelling checker/corrector for Basque based on two-level morphology", in *Proceedings of the Third Conf. on Applied Natural Language Processing (ANLP´92)*, Trento, 1992, 119-125.

[Agirre eta beste 94a] Agirre, E., Arregi, X., Arriola, J.M., Artola, X., Insausti, J.M.: *Euskararen Datu-Base Lexikala (EDBL)* (barne-txostena UPV/EHU/ LSI / TR 8-94), 1994.

[Agirre eta beste 94b] Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K.: "Conceptual Distance and Automatic Spelling Correction", in *Workshop on Comp. Linguistics for Speech and Handwriting Recognition*, Leeds University, 1994.

[Agirre eta beste 94c] Agirre, E., Arregi, X., Artola, X., Díaz de Ilarraza, A., Evrad, F., Sarasola, K.: "Sistema Diccionarial Multilingüe: aproximación funcional", in Boletín *SEPLN* 1, 313-333. Santiago de Compostela, 1994.

[Artola 93] Artola, X.: "HIZTSUA": Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception d'un système intelligent d'aide dictionnariale (SIAD)" Doktorego-tesia, EHU, 1993.

[Euskaltzaindia 85] Euskaltzaindia: *Euskal Gramatika: Lehen urratsak (I, II eta III)*. Euskaltzaindia, Bilbo 1985, 1990.

[Euskaltzaindia 91] Euskaltzaindia: *Euskaltzaindiaren Gomendioak eta Erabakiak (I eta II)*, 1992.

[Goenaga 80] Goenaga, Patxi: *Gramatika Bideetan*, Erein, Donostia, 1980.

[Gojenola 94] Gojenola, Koldo: "Aplicación de la relajación gradual de restricciones para la detección y corrección de errores sintácticos", in *SEPLN*, Córdoba, 1994.

[Karlsson eta beste 92] Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A.: *Constraint Grammar: A language-independent system for parsing unrestricted text* (eskuizkribua), Department of General Linguistics, University of Helsinki, 1992.

[Koskenniemi 83] Koskenniemi, K.: *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics. Publications 11, 1983.

[Díaz de Il. & Maritxalar 94] "An Icall System for studying the learning process", in *Computers in Applied Linguistics Conference*, Iowa State, USA, 1994.

[Zubizarreta 93] *Un modelo funcional para diálogos orientados por la tarea*, Doktorego-tesia, EHU, 1993.