Hizkuntz baliabideak Interneten

✗ **Proiektuaren izenburua:** Euskararen Datu Base Lexikalaren (EDBL) erabilera publikorako ingurunea.
✗ **Proiektuaren helburua:** Euskararen tratamendurako IXA taldearen zenbait produkturen erabilera Interneten zabaltzea.
✗ **Zuzendaria:** Xabier Artola Zubillaga.
✗ **Lan-taldea:** IXA taldea
E. Agirre, I. Aldezabal, I. Alegria, O. Ansa, X. Arregi, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, J.M. Intxausti, M. Lersundi, A. Maritxalar, M. Maritxalar, M. Oronoz, K. Sarasola, A. Soroa, R. Urizar eta M. Urkia.
✗ **Departamentua:** Lengoaia eta Sistema Informatikoak
✗ **Zentrua:** UPV-EHUko Informatika (Donostia)

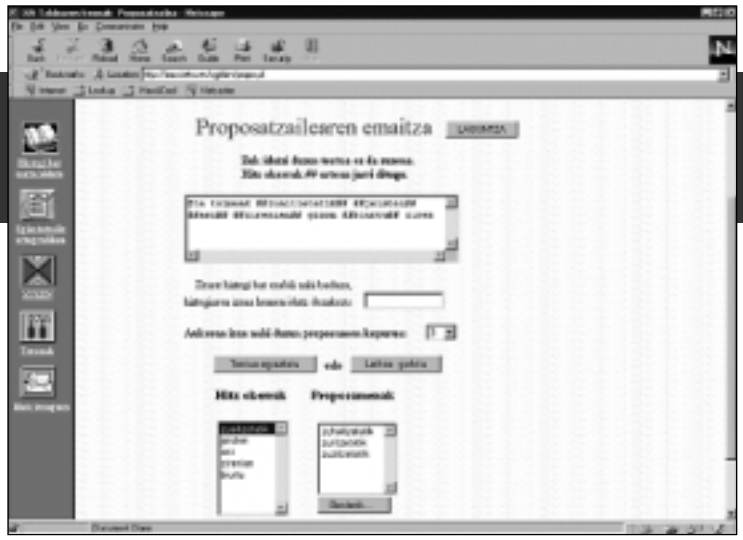
Ordenadorearen bitartez hizkuntzaren tratamendua egiten duten programak gero eta ugariagoak dira. Ordenadoreekin hizkuntza naturalen bitartez (euskaraz gure kasuan) komunikatu ahal izatea gero eta arruntagoa izango da. Bestetik, gizarte eleanitz honek hizkuntzen artean egin behar izaten dituen joan-etorriak leuntzeko ere aparteko lagun bihurtzen zaigu ordenadorea. Gainera, telekomunikazioetan gertatutako aurrerapen izugarriak (Internet fenomenoak batez ere) areagotu egin du hizkuntzaren tratamendu automatikoaren beharra. Izan ere, sarearen bidez informazio asko lor daiteke, baina ez da erraza behar dugun datu zehatz hori aurkitzea. Lan horretan tratamendu linguistikoa lagungarria baino ezinbestekoa da.

Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari *Lengoaia Naturalaren Prozesamendua* (LNP) esaten diogu. Hizkuntzaren inguruan industria berri osoa sortzen ari da, ordenadoreaz baliatuz hizkuntza tratatzea helburu duena. *Hizkuntzaren teknologiaz*, *hizkuntza-ingeniaritzaz*, hitz egiten da dagoeneko. Aplikazio-eremu nagusiak lau dira: i) Testu-edizioa edo testu-gestioa (ortografiaren eta estiloaren zuzentzaileak, testu eleanitzak sortu eta erabiltzeko laguntzak, hiztegi-kontsultak, ...); ii) Testu-masa handien tratamendu eta gestioa (kontzeptu-bilaketa, dokumentu-sailkapena, informazio-erauzketa eta testu-sorkuntza automatikoa); iii) Itzulpen automatikoa edo itzulpen lagundua, eta iv) Mintzoaren ezagutzea eta sorkuntza.

IXA taldean hamar urtez aritu izan gara arlo honetan, beti ere euskararen ikuspuntutik. UPV-EHUko Donostiako Informatika Fakultateko kideak eta UZEIkoak batuz 21 lagun gara guztira. Gure estrategia ez da inoiz sistema oso konplexua egitea izan, adibidez itzulpen-sistema egitea. Nahiago izan dugu helburu xume baina funtsezkoetatik hastea –adibidez morfologiatik, beste hizkuntzetarako arazo sinpleegitzat hartzen denetik– , eta bide horretan hizkuntz oinarri zabal eta sendoak eraikitzea. Geroago proiektu konplexuagoei ekin diegu –adibidez lematizazioari, sintaxia edo hiztegien erabilerari– baina aurretik eraikitako oinarri zabalaren gainean aritzeak denbora aurrezten digu eta sendotasuna ematen die produktu berriei. Gure hizkuntz baliabide horiek beste talde batzuentzat ere baliagarriak izan daitezkeenez, "*erakusketa elektronikoa*" zabaltzea erabaki genuen, eta horixe da, hain zuzen, artikulu honetan aurkezten dugun proiektuaren xedea. Proiektua Eusko Jaurlaritzako Unibertsitate eta Enpresaren arteko ikerketa-proiektuen 1997ko deialdian onartu zen (UE97/8 erreferentzia) eta 1998-99 urteetan burutuko da.

Epe ertainean Interneten kokatu nahi ditugun baliabideak hauek dira: datu-base lexikala, zuzentzaile ortografikoa, analizatzaile morfologikoa, lematizatzailea eta analizatzaile sintaktikoa. Baina oraingo lehen urrats honetan lehenengo hirurak bakarrik azalduko dira. Proiektua martxan ari da eta dagoeneko **http://ixa.si.ehu.es/tresnak** helbidean zuzentzaile ortografikoarekin probak egin daitezke (ikus itzazu artikulu hone-

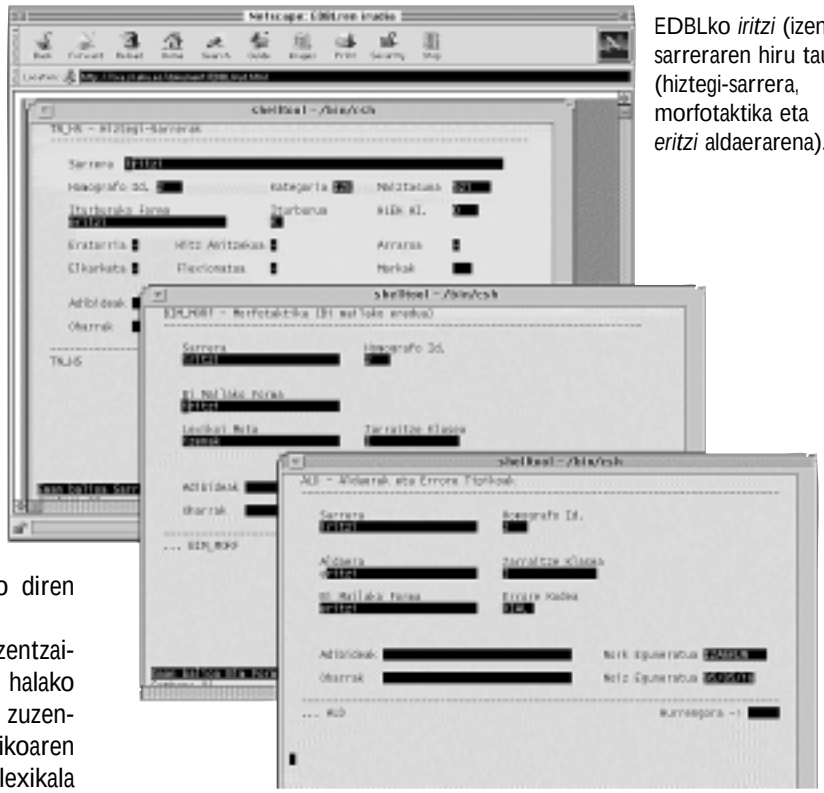
Taldeko aplikazioen web orria: http://ixa.si.ehu.es/tresnak/



Interneteko zuzentzailea testu hau zuzentzen: *Eta tximuak zuaitzetatik jeisten asi zirenian gizon biurtu ziren.*

tan bertan azaltzen diren ordenadore-pantailak edo ikusi zuzenean zeure ordenadorean). Proba egin zuzentzaileak ezagutzen ez dituen zure hitzak zeure hiztegi pertsonalean sartzen eta egiaztatu hitz horien deklinabideko beste formak ere ezagutuko dituela handik aurrera.

Bukatzeko proiektuaren izenean aipatzen den Euskararen Datu Base Lexikala (EDBL) zer den azalduko dugu. Datu-base lexikala lexikoaren biltegi erraldoia da. Hiztegi elektroniko moduko bat da, hizkuntzaren tratamendu automatikoari begira eraikia, eta, beraz, hizkuntzaren tratamendua automatizatu nahi horrek dituen eskakizunak kontuan harturik antolatua. Horrek, noski, lexikoaren antolakuntza gero zertarako erabiliko den kontuan hartuz egitea eskatzen du, eta lexiko-deskribapenaren sistematizazio bat: sarreren kategoria-sistema bateratu eta homogeneoa erabiltzea, kategoria bakoitzeko elementuak behar den bezala deskribatzeko beharrezko diren ezaugarriak zehaztea, etab.



EDBLko *iritzi* (izena) sarreraren hiru taula (hiztegi-sarrera, morfotaktika eta *eritzi* aldaerarena).

Euskararen kasuan, IXA taldean Xuxen ortografia-zuzentzailearen prestatze-lanari ekin genionean sortu zitzaigun halako lexiko-biltegiaren premia. Gorago esan bezala, baina, zuzentzaile hori oinarrizkoagoa zen analizatzaile morfologikoaren azpiproduktutzat hartzen genuen guk, eta datu-base lexikala ere ez genuen antolatu nahi izan zuzentzaile horretarako hiztegi edo hitz zerrenda soil gisa, etorkizunean euskararen tratamendu automatikoaren arloko beste edozein tresna edo aplikaziotarako oinarri lexikal sendo gisa baizik. Eta horrela sortu zen EDBL, Euskararen Datu Base Lexikala, harrezkero gure lanetarako oinarri lexikala izan dena, etengabe eguneratuz joan dena, eta gaur edo bihar komunitate zabalagoari bere ateak irekiko dizkiona, oinarriak prestatze-bide honetaz beste batzuk ere balia daitezen.

Datu-basea diseinatzerakoan garrantzi handia eman zitzaion, bada, etorkizunean izan ditzakeen hedapenak onartzeko behar bezain malgua izateari eta, bereziki, bertan jasoko zen informazio linguistikoa ahalik eta erarik neutralenean deskribatzeari, hau da, formalismo edo teoria linguistikoetatik ahalik eta erarik independenteenean deskribatzeari.

EDBLk gaur egun 70.000 sarrera inguru biltzen ditu, hiru atal nagusitan sailkatuta: hiztegi-sarrerak (izenak, adjektiboak, aditzak, etab.), adizkiak (aditz-forma jokatuak) eta morfema ez-independenteak (atzizki, aurrizki, etab.). Sarrera-kategoria bakoitzeko aldez aurretik definiturik dauden ezaugarri edo atributuak erregistratzen dira, eta kasu guztietan, lehen aipatu bezala, sarrerari dagokion morfologia ere deskribatzen da (informazio morfotaktikoa), horretarako morfologia konputazionalean asko erabiltzen den bi mailatako formalismoaz baliatuz.

EDBL egun datu-baseen kudeaketarako sistema komertzial baten pean dago eta halako sistemek ohikoak dituzten erraztasunak eskaintzen dizkio hizkuntzalariari –hizkuntzalariak baitira bere erabiltzaile nagusiak–: interfaze atsegina lanerako, informazioa egunean mantendu eta berorren kontsistentzia bermatzeko erraztasunak, behar diren aplikazioetarako informazioa behar bezala iragazteko aukerak, eta abar. Euskararen baturatze-bidean izandako azken gertakariak –Euskaltzaindiaren erabakiak, batik bat– eguneratuta mantentzeko ere ezinbesteko tresna bihurtu da datu-basea, eta etorkizunean EDBLk bete dezakeen lan inportanteetariko bat izan daiteke azken erabakien berri emango duen tresna izatea.

IXA taldea