EUSKARAREN DATU-BASE LEXIKALA (EDBL)

1. Sarrera gisa.

Txosten honen helburua Euskararen Datu-Base Lexikalaren (EDBL) diseinu berria aurkeztea da. Jakina denez, datu-base hori XUXEN egiaztatzaile/zuzentzaile ortografikoa garatzeko ezinbesteko euskarri lexikal gisa sortu zen. Hala ere, hasiera-hasieratik helburu orokorrago batekin jaioa —zuzentzailearekin batera analizatzaile morfologiko orokorra ere garatu baitzen[1]—, EDBL euskararen tratamendu automatikoaren bidean eginiko lehen tresna orokorra den horren euskarri lexikala dugu.

Hasiera batean egindako datu-basea orain arte baliagarri izan bazaigu ere, akatsak eta hutsuneak dituena ezin ukatuko dugu. Batik bat euskararen tratamendu automatikoan bestelako lanak hartu nahi ditugunean, eta guztietan oinarri lexikal sendoaren premia ezinbesteko dugunean. Hori dela eta ekin genion iaz (92-93 ikasturtean) EDBLren birdiseinatze lanari.

EDBLren birdiseinuari begira orain arte eman ditugun urratsen segida laburbilduz, bi fase bereiz daitezke:

- Lehen fase batean (1993ko uztaila), EDBL "zaharretik" abiatuz EDBL osoago eta orokorrago bat proposatu zen. Lehenengo saioan egindako diseinuari, alabaina, oinarrizko akats bat bilatu zitzaion, alegia, datu-basearen antolaketa bi mailatako ereduaren menpe geratzen zela[2], "menpeegi" aukeran, EDBLrentzat nahi zen neutraltasuna kontutan hartuz behintzat. Proposamen hartan sarreren gako nagusitzat maila lexikaleko forma hartzen genuen, bere hautapen-markak eta guzti.

  Lehen fase honetako bigarren saioan irtenbidea bilatu zitzaion arazo horri,bi mailatako ereduari zegokiona "bigarren maila" batean utziz, sarrerak unibokoki identifikatzeko homografo-zenbakia sartzea proposatu baitzen. Hiztegi-sarreretan ez ezik, bestelako morfema eta adizkietan ere berdin jokatzea erabaki genuen, alegia, idazkera bereko sarrerak "homografo-zenbaki" baten bitartez bereiztea, artifizialxeagoa izanik ere irtenbide dotore eta eraginkorra zelakoan.

---

[1] Hobeto esanda, zuzentzaile ortografikoa analizatzaile morfologikoaren azpiproduktua da.

[2] Bi mailatako eredua da analizatzaile morfologikoaren garapenean euskararen morfologia deskribatzeko erabilitako formalismoa (Koskenniemi, 83).

Aurreko bi bilera horietan beste eztabaidagai funtsezko bat ere izan genuen; labur esanda, kontua zen ea zer zen komenigarriago, edo (1) taula gutxitan era askotako informazioa biltzea (horretarako kategoria desberdineko sarrerek behar zuten orotariko informazioa taula erraldoi batzuetan bilduz, eremu asko eta asko, kasu gehienetan, hutsik utziz), edo, bestela, kategoria desberdinetako sarrerek oso informazio-eremu desberdinak behar dituztenez, (2) kategoria bakoitzerako taula bereziak diseinatu eta taula horietako bakoitzean kategoria horretako sarrerei legokiekeen informazioa errepresentatu. Horiek horrela, interfazea diseinatzeko erosotasunari begiratuz, batik bat, lehenengo bideari jarraitzea erabaki genuen.

- Birdiseinatze lan honetako bigarren fasea berriz, txosten honek hartzen du bere baitan.

  Proposamen honetan, arazoa ostera azterturik, eta beste datu-base lexikal batzuk ikusirik (Pin-Ngern *et al.*, 93), aurreko bi bide horietako bigarrena hartzea proposatu nahi dugu: kategoria bakoitzerako taula bereziak diseinatu eta taula horietako bakoitzean kategoria horretako sarrerei legokiekeen informazioa errepresentatu. Bide horren abantailarik nagusiena, gure iritziz, EDBL "irekia" uzten duela da, etorkizunean erantsi nahi izango dizkiogun informazio-eremu berriak aise eransteko moduan, alegia. Inongo dudarik gabe, datu-basearen irekitasun horren alde egin behar apustu, batez ere, EDBL hizkuntzaren tratamendu automatikoan hainbat eta hainbat lanen euskarri izatea nahi baldin badugu. Egun ezin baita esan zer-nolako informazioa beharko den bihar egin nahiko dugun ez dakit zein lanetan, eta lehenengo bidea hautatuz gero, datu-basearen eskema aldagaitz eta estuegi bat izango genukeelakoan gaude.

Iazko bi saio horien ondoren etorri da aurten (94ko urtarrila) mintegian landutako beste gai bat, EUSLEM lematizatzaile/etiketatzailerako behar diren kategoria-etiketei buruzkoa alegia. Kategoria-etiketa horien azterketak lexikoko elementuen kategoria nagusien —eta azpikategorien— errebisioa ekarri du berarekin eta orain arte erabilitako kategoria-sistema zuzendu eta aberastu egin dela esan dezakegu. Horrek ere eragina izan du, noski, orain hemen aurkeztera goazen lan honetan.

Sarrera honen ondoren bada, ikus dezagun EDBLren diseinu berri hau zertan den. Esan gabe doa txosten honetako guztia proposamen mailan hartu beharra dagoela eta erabat zabalik dagoela egin nahi diren hobekuntzak eta aldaketak onartzeko.