DEPENDENTZIA-EREDUAN OINARRITUTAKO BALIABIDE SINTAKTIKOAK: ZUHAITZ-BANKUA ETA GRAMATIKA KONPUTAZIONALA

Tesiaren egilea: Mª Jesús Aranzabe Urruzola
Unibertsitatea: Euskal Herriko Unibertsitatea
Saila: Euskal Filologia
Tesi-zuzendaria: Jose Mari Arriola eta Arantza Díaz de Ilarraza
Tesiaren laburpena:

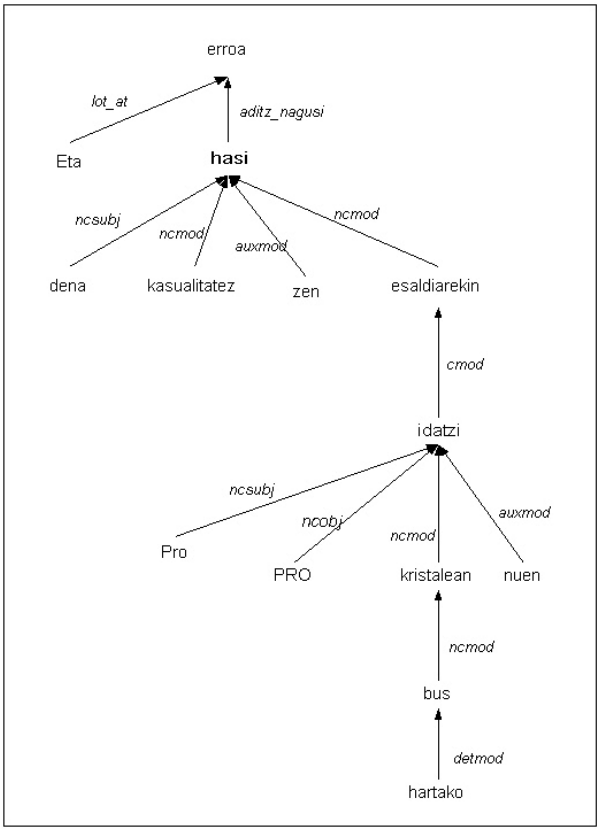
Tesi-lan hau Hizkuntzaren Prozesamenduaren barruan kokatzen da; zehatzago esateko, hizkuntza ulertu ahal izateko behar-beharrezko den sintaxi konputazionala izan du langai.

Bere helburu nagusia euskarazko testu errealak aztertzeko baliabide sintaktikoak garatzea da; zuhaitz-bankua (edo *treebank*a) eraikitzea eta gramatika konputazionala definitzea, alegia. Zeregin horretan kontuan hartu dira IXA taldearen ikerkuntza-bideak (ikusi http://ixa.si.ehu.es/Ixa), urratuak zein urratzen ari direnak, batetik, aurkeztuko den lan hau horietan integratzeko eta, bestetik, berau gauzatzen laguntzeko.

Testu errealen analisi sintaktikoa gauzatzeko lehendabiziko lana, deskribapen formala eta aplikazio informatikorako egokia den eredua aukeratzea da. Behin eredu desberdinak aztertuta eta erdal hizkuntzetan egin diren *treebank*ak gainbegiratu ondoren, Dependentzia Gramatikaren ereduari (Tesnière, 1959) jarraitzea erabaki da. Eredu horretan, esaldiko hitzak binaka lotuz esaldiaren *dependentzia-zuhaitza* lortzen da (1. irudia). Zuhaitz sintaktiko hauetan, batetik, adabegietan dauden hitzen arteko gobernatzaile/mendeko erlazioak irudikatzen dira, eta bestetik, bi hitzen arteko loturan mendekoak betetzen duen funtzio sintaktikoa adierazten da dependentzia-etiketen bidez.

Eredua aukeratu ondoren, zuhaitz-bankua eraikitzeko metodologia definitu eta 300.000 hitzeko *Euskararen Prozesamendurako Erreferentzia Corpusa* (EPEC) sintaktikoki etiketatu da eskuz.

Mota honetako corpusen garrantzia ukaezina da, hizkuntza teknologiaren produktuak ikertzeko eta garatzeko oinarri enpirikoak jartzen dituelako eskura, besteak beste. Ikuspegi teorikotik, corpus horretan zehaztu den informazio sintaktikoa oso lagungarria izango da hizkuntzari buruzko ikerketa linguistikoak egiteko, eta ikuspegi konputazionaletik, berriz, corpus etiketatuaren bitartez tresna informatikoak garatu, ebaluatu eta informazio linguistikoa automatikoki eskuratu ahal izango da.



1. irudia: ***Eta dena kasualitatez hasi zen, bus hartako kristalean idatzi nuen esaldiarekin*** **esaldiaren dependentzia-zuhaitza.**

Sintaktikoki etiketatutako corpus honek erabilera bat baino gehiago izan ditu dagoeneko; hala, metodo edo sistema berriak probatzeko urtero antolatzen den *Conference on Computational Natural Language Learning* (CoNLL) lehiaketan erabili da (Nivre *et al.*, 2007), eta baita estatistikan oinarritutako euskararen analizatzaile sintaktikoa garatzeko ere (Bengoetxea eta Gojenola, 2007). Tesi-lan honetan, hizkuntza-ezagutza hori Euskararen Dependentzia Gramatika Konputazionalak (EDGK-I eta EDGK-II) definitzeko eta ebaluatzeko erabili da. Gramatika horiek analizatzaile sintaktikoaren bitartez aplikatu ondoren, corpus edo testuetako esaldien analisi sintaktiko osoak lortzen dira; hau da, hitz isolatuez osatutako sekuentzietan elkarrekin lotuta dauden egitura sintaktikoak (sintagmak, perpausak, esaldiak…) ezagutzen dira. Horien aplikazioa bi urratsetan egiten da. Lehen urratsean, EDGK-I gramatika baliatuta corpuseko hitz bakoitzari dependentzia-etiketa bat esleitzen zaio. Eta bigarren urratsean, EDGK-IIren bitartez, esleitutako dependentzia-etiketa horien arteko loturak (gobernatzailearen eta mendekoaren artekoak) esplizitu egiten dira.

Ondorio nagusi gisa, lan honetan analisi sintaktiko partzialetik analisi sintaktiko osorako pausoa eman da eta Dependentzia Gramatikan oinarrituta euskararen sintaxiaren lehen formalizazioa egin da.