ALDAERA LINGUISTIKOEN NORMALIZAZIOA INFERENTZIA FONOLOGIKOA ETA MORFOLOGIKOA ERABILIZ

Tesiaren egilea: Izaskun Etxeberria Uztarroz
Unibertsitatea: Euskal Herriko Unibertsitatea (UPV/EHU)
Saila: Lengoaia eta Sistema Informatikoak Saila
Tesi-zuzendaria: Iñaki Martínez de Albeniz Ezpeleta eta Elena Casado Aparicio
Tesiaren laburpena: Iñaki Alegria eta Montse Maritxalar

Hizkuntzaren azterketa eta prozesamenduaren barnean, tesi-lan hau testu ez-estandarren ikertze-arloan kokatzen da, euskarazko testu ez-estandarren arloan, zehazki. Oro har, testu estandarrekin alderatuta, testu ez-estandarrek ezaugarri bereziak dituzte maila lexikoan, morfologikoan edota fonologikoan, eta hala, haien prozesaketa erronka bat da.

Euskararen kasuan, oso jatorri eta garai ezberdineko testuak dira ez-estandarrak. Alde batetik, ez-estandarrak dira euskararen estandarizazio-prozesua baino lehen idatzitako testuak (XVI. mendetik XX. mende erdira arte sortutakoak). Beste aldetik, estandarizazio-prozesuaren ondoren idatzitako guztia ez da estandarra izan, noski, dialekto aberats eta ugariak baititu euskarak, eta horiek ere erabili dira eta erabiltzen dira egun. Halaber, testu ez-estandar ugari aurkitzen dugu gaur egun sare sozialetan; askotan sailkatzeko zailak dira —ez dago argi dialektalak diren edo beste fenomeno batzuen adierazleak diren—, baina, dena dela, ez-estandartzat jo behar dira horiek ere.

Jatorria alde batera utzita, testu ez-estandarren prozesatzeak, oro har, arazo berberarekin egiten du topo beti: hizkuntza prozesatzeko tresna gehienak hizkuntza estandarretan idatzitako testuak prozesatzeko garatu dira, eta testu ez-estandarrekin erabiltzen direnean, asko jaisten da haien errendimendua.

Halako testuak prozesatzeko interesa, ordea, asko zabaldu da azken urteetan. Adibide bat liburu digitalen eskutik dator. Gaur egun liburutegi digital ugari daude Internet bitartez atzigarri, publiko zabalari aukera berriak eskaintzen dizkiotenak: orain arte adituek soilik kontsulta zitzaketen zenbait dokumentu preziatu eta urri guztion eskura daude liburutegi horien bitartez. Dokumentu horietako asko aspaldikoak direnez, ez daude idatzita hizkuntza estandarraren arabera, eta bertan kontsultak egitea ez da egungo dokumentuetan bezain erraza.

Liburutegi digitalez gain, Humanitate Digitalen arlo berria dago. Informazio digitala prozesatzeko baliabideak humanitateen ikerkuntzaren eskura jartzea da arlo horren jomuga, eta horrek zera eskatzen du, humanitateetako tradiziozko metodo kualitatiboak eta egungo aplikazioak —informazio-berreskuratzea, informazio-erauzketa, datu-meatzaritza eta abar— nolabait biltzea. Hala, humanitateentzat interesekoak diren testuen artean, testu historiko ugari daude, hau da, testu ez-estandar ugari, eta horiek prozesatu beharra dago.

Adibide gehiago badaude ere, gakoa da testu ez-estandarrak normalizatuz gero aukera dagoela hizkuntza prozesatzeko tresnak aplikatzeko testu horietan, eta, beraz, funtsezkoa da normalizazio-prozesu hori ahalik eta modurik eraginkorrenean betetzea. Tesi-lan honetan ikasketa automatikoan oinarritzen diren metodoak proposatzen dira euskarazko testu ez-estandarretan normalizazioaren ataza ebazteko. Gure helburu nagusia da morfofonologia konputazionaleko tresnak erabiltzea metodo bat ikasteko eta gai izateko euskarazko aldaerei —diakronikoei zein dialektalei— dagozkien forma estandarrak automatikoki esleitzeko. Horrekin batera, gure proposamenak lortzen dituen emaitzak konparatzen dira beste ikerketa batzuek lortzen dituztenekin, horrela metodoen egokitasuna aztertzeko. Konparazio hori egiteko gaztelaniazko zein eslovenierazko corpusak erabili dira, beste zenbait ikertzaileren lankidetza baliatuz.