MATXIN. ERREGELETAN OINARRITUTAKO ITZULPEN AUTOMATIKOKO SISTEMA BATEN ERAIKUNTZA ESTALDURA HANDIKO BALIABIDE LINGUISTIKOAK BERRERABILIZ

Tesiaren egilea: Aingeru Mayor Martinez
Unibertsitatea: Euskal Herriko Unibertsitatea
Saila: Lengoaia eta Sistema Informatikoak
Tesi-zuzendaria: Kepa Sarasola eta Arantza Díaz de Ilarraza
Tesiaren laburpena:

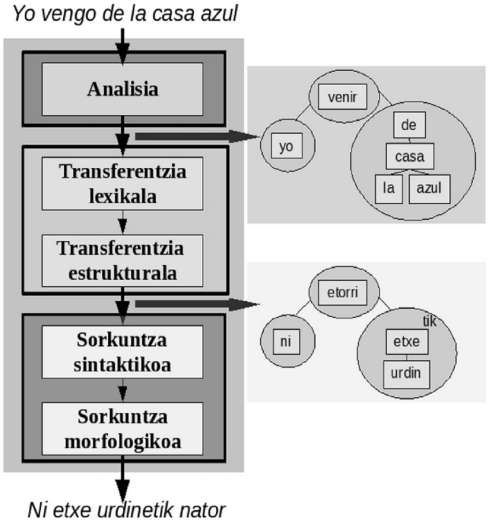
Itzulpengintza automatikoa (IA) mundu globalizatuan gero eta garrantzi handiagoa hartzen ari den garai hauetan, euskararako publikoki erabilgarria den lehenengo itzulpen-sistema aurkezten dugu: *Matxin*. Ikerketa-lan hau IXA taldearen barruan garatu da.

Azken hamarkadan IAren arloan arrakasta handia izan duten hurbilpen estatistikoek zailtasunak dituzte euskararekin lan egiteko: batetik, euskara hizkuntza eranskaria delako, eta, bestetik, euskararako dauden corpusen tamaina mugatua delako. Gure apustua erregeletan oinarritutako sistema tradizional bat sortzea izan da, aurreikusten baitugu estrategia honek pisu handia izango duela etorkizunean euskararako eraikiko diren sistema hibridoetan

Matxin erabilera orokorreko sistema da eta hiru fasetan egiten du itzulpena: analisia, transferentzia eta sorkuntza.

Analisi-fasean abiapuntu-testuaren analisi sintaktiko partziala *Freeling* software libreko paketea erabiliz lortu ondoren, guk sortutako modulu batek beharrezkoa izango den hitzen eta sintagmen arteko mendekotasun-informazioa ematen du.

Transferentzia lexikalean estaldura handiko hainbat hiztegi berrerabiliz eraiki dugun lexikoi elebiduna kontsultatzen da. Transferentzia estrukturalean, besteak beste, postposizioen desanbiguazioa (adibidez, nola itzuli *de la casa*? *etxearen*, *etxeko* ala *etxetik?*) eta aditz-kateen transferentzia burutzen da.



Sorkuntza sintaktikoan hitzak eta sintagmak ordenatzen dira eta sorkuntza morfologikoan postposizioen informazioa sintagmako azken hitzari gehitzen zaio, hitzaren forma lortuz *Morfeus* prozesadore morfologikoarekin.

Hainbat tresna eta baliabide linguistiko berrerabili ditugunez, bereziki landu behar izan dugu sistemako datuen, erregelen eta moduluen arteko datu-fluxuaren formatuen estandarizazioa.

Sistemaren arkitektura abiapuntu-hizkuntzatik independetea izateko diseinatu dugu eta inplementatu dugun *Matxin1.0* prototipoak espainieratik euskarara itzultzen du. Prototipo hau *Opentrad* proiektuko web orrian proba daiteke (http://www.opentrad.org) eta kode irekiko software libre bezala banatzen da (http://matxin.sourceforge.net). Sistema ingelesetik euskarara itzultzeko ere egokitzen hasiak gara.

Sistemaren ebaluaziorako *edizio-distantzia* neurria erabili dugu. Emaitza % 40 izan da, hau da, 10 hitzetatik 4 aldatu behar dira sistemak emandako irteeratik itzulpen onargarri bat lortzeko. Atazaren konplexutasuna zein den kontuan hartuta, emaitza positiboki baloratzen dugu, beste hizkuntza bateko testuak ulertzeko baliogarria delako. Gainera, metodo estatistikoetan oinarritutako Dublin City Unibertsitateko *Matrex* espainiera-euskara itzulpen-sistemarekin konparatu dugu eta gure emaitzak hobeak dira (% 40 vs % 60).

Aurrera begira, eraiki dugun sistema hobetzeko eta teknika eta estrategia berriak aztertu eta integratzeko lan asko dugu egiteko. Umiltasunez aitortzen dugu guk egindako lana hasiera baino ez dela; harrotasunez onartzen dugu gure ekarpenak bide asko zabaltzen dituela.