# AZALEKO SINTAXIAREN TRATAMENDUA IKASKETA AUTOMATIKOKO TEKNIKEN BIDEZ: EUSKARAKO KATEEN ETA PERPAUSEN IDENTIFIKAZIOA ETA BERE ERABILERA KOMA-ZUZENTZAILE BATEAN

**Tesiaren egilea:** Bertol Arrieta Kortajarena
**Unibertsitatea:** Euskal Herriko Unibertsitatea
**Saila:** Lengoaia eta sistema informatikoak
**Tesi-zuzendaria:** Iñaki Alegria Loinaz eta Arantza Díaz de Ilarraza Sánchez
**Tesiaren laburpena:**

XUXEN ortografia-zuzentzailearen arrakastaren ondoren eta IXA taldean Hizkuntzaren Prozesamenduan urtetan egindako lanari jarraiki, XUXENg euskarako gramatika- eta estilo-zuzentzailea garatzeko aurrerapausoak aurkeztu ditugu tesilan honetan. Zehazki, koma-zuzentzaile bat garatu dugu, etorkizunean gramatika- eta estilo-zuzentzailean txertatzeko asmoz.

Erroreen detekziorako bi hurbilpen erabili izan ohi dira hizkuntzalaritza konputazionalean: hizkuntza-ezagutzan oinarritutakoa, batetik, eta corpusetan oinarritutakoa, bestetik. Ikasketa automatikoko teknikak bigarren hurbilpenaren baitan sailkatzen dira. Izan ere, corpus handietan duen informazioa —informazio morfosintaktikoa, gure kasuan— baliatuz nolabait *ikasten* saiatzen da makina, ikusi gabeko testuetan gerora erabakiak hartzeko. Teknika hauek baliatu ditugu, batez ere, tesi-lan honetan, nahiz eta hurbilpen bateko eta besteko teknikak uztartzera ere jo dugun.

Bestalde, azaleko sintaxian ere egin dira hainbat aurrerapen. Zehazki, euskarako kateen eta perpausen identifikatzaile automatiko lehiakorrak sortu ditugu, lehendik bazirenak hobetuz. Horretarako, ikasketa automatikoko tekninak erabili eta hizkuntza-ezagutzan oinarritutakoekin uztartu ditugu. Sintagmak eta aditzkateak jo ditugu katetzat, oro har. Kateen eta perpausen identifikazioa garrantzitsua da Hizkuntzaren Prozesamenduko hainbat arlotan (itzulpen automatikoan, kasu), baina behar-beharrezkoak dira komen zuzenketarako: oro har, ez da komarik joango kate baten baitan, eta perpaus-mugetan komak jarri behar izaten dira kasu anitzetan.

Puntuazio-arauak hizkuntza askotan ez daude guztiz zehaztuta, eta euskaraz are gutxiago. Badira fidagarritzat jotzen diren puntuazio-markak (puntua, galdera-marka eta harridura-marka), baina koma ez fidagarrien artean dago: gehien erabiltzen dena da, baita modu zabalenean ere, baina bere erabilera ez da oso estandarra, eta gutxien arautua dagoena da. Koma-zuzentzaile baten garrantzia, beraz, ukaezina da. Gainera, Hizkuntzaren Prozesamenduan, komak zuzen izateak analisi sintaktiko hobea lortzen lagundu dezake, besteak beste.

Koma-zuzentzailea sortzeko, hainbat urrats eman behar izan ditugu, baina komaren arauen formalizazioa egitea izan da lehendabizikoa, hainbat hizkuntzalari adituren laguntzarekin. Gero, IXA taldeak sortutako analizatzaile sintaktikoak eta guk landutako kate- eta perpaus-identifikatzaileek emandako informazioa baliatuz eta ikasketa automatikoko tekniketan oinarrituz, hitz bakoitzaren ondoren koma datorren ala ez erabakitzen duen sailkatzailea sortu dugu. Hauxe izan da koma-zuzentzailearen hazia.

Lortu ditugun emaitzek gure tresna baliagarria egiten dutela uste dugu, baina tentuz erabiltzekoa izango da, doitasun perfektutik urrun baikaude gaur-gaurkoz.