Errore sintaktikoak automatikoki detektatzen eta zuzentzen

Ordenagailuaren aurrean euskaraz eta euskararekin lan egiten dugunok, Xuxen euskarako zuzentzaile ortografikoa maiz erabili izan dugu. Testu-editoreren bat erabiltzen ari garenean, Xuxenek "xuxen" idatzi ez ditugun hitzetako erroreak salatzen dizkigu, baita ordain zuzenak eman ere. Baina hori da, hain zuzen ere, zuzentzaile ortografikoen muga: hitza.

Muga hori gainditzeko asmoz, erroreen detekzio eta zuzenketa automatikoaren alorrean hitzetatik haratago joan nahi izan dugu IXA taldean[1], eta sintagmetan zein esaldietan gertatzen diren hainbat errore landu ditugu. Sintaxi-erroreen detekzio automatikorako erregelak garatuta, horiek bi eremutan erabiltzeko asmoa dugu: i) XUXENg euskararako gramatika-zuzentzailean, eta ii) ordenagailuek lagundutako hizkuntzen irakaskuntzan.

Sintagma-mailako errore batzuk automatikoki lantzea erraza da; adibidez, "*guzti hori" egitura erroredunean akatsa salatzeko, hitzen ordena kontuan hartzea besterik ez da egin behar. Beste batzuetan, ordea, testu-zati batean errorea dagoela esatea ez da lan erraza izaten, eta gainera, informazio asko eta askotarikoa behar izaten da egitura zuzenak egitura erroredunetatik bereizteko. Bereizketa are zailagoa da erroreak konputazionalki lantzen ditugunean, ordenagailuei pertsonek dugun munduari buruzko ezagutza falta baitzaie. Esaterako, adibide hau erroreduna dela esango genuke gizakiok, dudarik gabe:

*[**Lagunak artean**] erosi diote.

Ordenagailuak, ordea, zaila du aurreko adibidea ondorengo esaldi zuzenetatik bereiztea:

[Lagunak artean] daude txantxangorriaren habia aztertzen.

[Lagunak artean] erosi gabe zioten oparia.

Aipatu berri ditugun moduko postposizio-lokuzioetan gertatzen diren erroreak automatikoki azalarazi ditugu, besteak beste, nire tesi-lanean. Alde batetik, esaldiko testuinguru lokalari (ondoz ondoko bospasei hitzei) erreparatuz detekta daitezkeen erroreak detektatu eta zuzendu ditugu. Eta bestetik, esaldiaren eremura helduta, subjektuaren, objektuaren eta zehar-objektuaren eta aditzaren arteko komunztadura-ezak landu ditugu. Komunztadura-erroreak lantzeko esaldiaren analisi-zuhaitza aztertzeak erroreen detekzio-lana izugarri arintzen du, eta horretarako, Saroi tresna sortu dugu. Has gaitezen bada, urratsez urrats erroreak eta horiek detektatzeko erabilitako teknikak azaltzen.

Esaldiko ordena linealean ondoz ondoko hitzetan gertatzen diren erroreak aztertzen hasiko gara. Egun, ohikoa da "2009-ko irailaren 28" moduko data okerrak idatzita ikustea. Data-ko elementuak banan-banan hartuta, hitzak zuzenak dira, Xuxen-ek ez lizkiguke azpimarratuko, baina ez al du marratxorik gabe eta "irailaren 28a" behar? Era honetako egitura erroredunak, hilabetearen eta egunaren arteko komunztadura aztertzen dutenak adibidez, markatu eta zuzendu ditugu. Horretarako, erroreetan agertzen diren egitura okerrak gramatika batean deskribatu ditugu, eta gero, *Xerox Finite State Tool* (XFST) izeneko tresna erabiliz, datekin osatutako zerrenda batean probatu ditugu.

Testuinguru mugatukoak dira ere, postposizio-lokuzioak. Egiturok lantzea, datak lantzea baino zailagoa egin zaigu, semantikaren egitekoa oso garrantzitsua baita postposizio-lokuzioetan. "Arte", "aurre", "bitarte", "buruz" eta "zehar" hitzak osagai beregaintzat dituzten postposizio-lokuzioak aukeratu ditugu, maiz gaizki erabiltzen direlako. Tratamendu konputazionalari dagokionez, alde handia dago aukeratutako postposizioen artean. Adibidez, "zehar" lantzeko ez dugu inolako arazorik izan (ohiko akatsa da genitiboan edo deklinatu gabe erabiltzea, "*basoaren zehar" adibidean kasu), eta gainera, oso emaitza onak lortu ditugu. "Arte" eta "buruz", ordea, buruhauste ugariren iturri izan ditugu. Egiturok oso anbiguoak dira, eta egitura zuzenak erroreduntzat jo izana, maiz gertatu zaigu. "Arte" postposizioaren konplexutasuna ebazteko, batzuetan ezaugarri semantikoak ere erabili behar izan ditugu, denborari edo lekuari buruz ari garen ezagutzeko. Adibidez, "gero arte" esan dezakegu, baina ez, "*etxea arte" ("etxeraino" behar du). Erroreen deskribapen osoa erregeletan kodetu dugu, eta sortutako gramatika *Constraint Grammar* tresna erabiliz testuei aplikatu diegu.

Komunztadura-erroreak, aurrekoak ez bezala, testuinguru zabalekoak dira eta mota honetako erroreak detektatzeko esaldiko ordena lineala erabiltzea ez da egokia. Har dezagun esaldi bat, hainbat ordenatan idatzita:

1. *Zentral nuklearrak zakar erradiaktiboa eratzen dute.
2. *Zakar erradiaktiboa eratzen dute zentral nuklearrak.
3. *Zentral nuklearrak eratzen dute zakar erradiaktiboa.

Hiru adibide horietan, errorea bera da: subjektuak ("zentral nuklearrak") eta aditz laguntzaileak ("dute") ez dute komunztadura ongi egiten kasuan edota numeroan. Orain arte aipatutako teknikak erabiliz gero, hiru erregela beharko genituzke ordena ezberdineko hiru esaldiotan errore bera detektatzeko. Esaldiaren egiturazko ordena, edo analisi-zuhaitza erabiliz gero, ordea, erregela bakarra behar dugu (hiru esaldiek analisi-zuhaitz bera dute). Beraz, egokiena, esaldiari dagokion analisi-zuhaitza eraiki, eta zuhaitz horretan erregela baten bidez egitura erroreduna topatzea da. Horretarako sortu genuen, *Saroi* tresna (Sintaxi ARboletan Oinarrizko bilaketak). Hasiera batean *Saroi*ren helburua erroreen detekzioa bazen ere, laster konturatu ginen tresna orokorragoa izan zitekeela. *Saroi*ren helburu nagusia esaldi erreal bati dagozkion mendekotasun-zuhaitzetan informazioa bilatzea da. Euskaraz, dagoeneko sortua dugu eskuz etiketatutako mendekotasun-zuhaitzen banku edo *treebank* bat. *Saroi* baliatuz edozein ikertzailek aukera izango du euskaraz erabiltzen diren egiturei buruzko kontsultak zuhaitz-bankuan egiteko.

Maite Oronoz
IXA taldeko kidea eta EHUko irakaslea

[1] http://ixa.si.ehu.es