EusCrawl izeneko euskal corpus librea sortu du EHUko IXA taldeak

Paul Picado

Guztira, 12,5 milioi dokumentuk eta 423 milioi hitzek osatzen dute corpusa

EusCrawl izeneko euskal corpus librea sortu du Euskal Herriko Unibertsitateko IXA ikerketa-taldeak. Taldeak hizkuntzaren tratamendu automatikoan lan egiten du, eta oraingoan, 12,5 milioi dokumentu eta 423 milioi hitzekin corpusa osatu du. Hori guztia euskarazko 33 webguneetatik atera dute, horien artean, Wikipedia, Berria, Argia, Hitza eta Bilbo Hiria Irratia.

Webgunean dioten moduan, corpusa osatzeko, eskuz aukeratutako Interneteko hainbat webguneetatik dokumentuak xurgatu dituzte, eta hortik dator, hain zuzen, crawl ingelesezko hitza, xurgatu esan nahi duena.

IXA taldeak dioenez, corpus bat osatzeko testu kopuru handia behar da. Ingelesaren kasuan, esaterako, erraza da, testu ugari daudelako ingelesez idatzita Interneten. Euskaraz, berriz, baliabide gutxi daude, eta horrenbestez, testu kopuru handia biltzea zaila da.

EusCrawl corpusaren aurretik beste batzuk ere bildu dira aurretik. Orduan, zer abantaila dauka EusCrawl-ek? Bada, besteak edonork kontsulta ditzake, baina EusCrawl osorik deskargatu eta berrerabiltzeko aukera dago. Horrenbestez, benetako azterketa linguistikoa eta ikerketa egiteko aukera ere ematen du. Horri esker, esaterako, BigScience proiektuan erabiliko dute, hizkuntza-eredu eleaniztun

eta erraldoi librea eraikitzeko.

IXA ikerketa-taldeko bost kidek lan egin dute proiektu honetan, hala nola, Mikel Artetxe informatikariak, Itziar Aldabek, Rodrigo Agerrik, Olatz Perez de Viñasprek eta Aitor Soroak, eta Meta enpresak ere parte hartu du. Corpusa, xede horretarako sortu duten EusCrawl webgunean bertan eskura daiteke, hemen.

- Guerra en Ucrania
- Athletic
- Zurekin
- Bizkaia Dmoda
- Antropía