EUSKARAZKO DENBORA-EGITUREN AZTERKETA ETA CORPUSAREN SORRERA /ANALYSIS OF BASQUE TEMPORAL CONSTRUCTIONS AND THE CREATION OF A CORPUS

Tesiaren egilea: Begoña Altuna Díaz
Unibertsitatea: Euskal Herriko Unibertsitatea (UPV/EHU)
Saila: Euskal Hizkuntza eta Komunikazioa Saila
Tesi-zuzendaria: Arantza Díaz de Ilarraza eta María Jesús Aranzabe
Tesiaren laburpena:

Denbora-informazioak testuko informazioa ardatz kronologikoan kokatzen laguntzen du. Hau da, denbora-informazioak *zer noiz* gertatu den adierazten du. Gizakiok etengabe hitz egiten dugu gertatutakoaz edo etorkizuneko planez eta ekintzak eta egoerak —gertaerak— aise kokatzen ditugu denboran. Tresna automatikoentzat, ordea, ez da hain erraza gertaerak denboraren arabera antolatzea. Tesi-lan hau Hizkuntzaren Prozesamenduan (HP) kokatzen da eta helburu nagusitzat du euskarazko testuetako denbora-informazioaren ulermen automatikorako oinarriak ezartzea.

Euskarazko denbora-informazioa HPn erabiltzeko jarraitu behar diren urratsak deskribatu ditugu tesi-lanean. Urrats horiek definitzeko, denbora-informazioaren prozesamenduan egindako beste lanak aztertu ditugu eta euskararen prozesamendurako baliabide eta ildo egokienak aukeratu ditugu. Definitutako urratsak hauek dira:

1. Euskaraz denbora-informazioa zein elementuk adierazten duen aztertu dugu. Zehazki, gertaerek, denbora-adierazpenek eta horien arteko erlazioek (denbora-erlazioak, aspektu-erlazioak eta mendekotasun-erlazioak) zein forma hartzen duten aztertu dugu. Era berean, horietan denbora-informazioa zein ezaugarriren bidez adierazten den identifikatu dugu.

2. Euskarazko denbora-informazioa kodetzeko EusTimeML markaketa-lengoaia eta testuak etiketatzeko gidalerroak definitu ditugu. EusTimeMLren bidez, gertaerak, denbora-adierazpenak, erlazioak eta denbora-erlazioak esplizitu egiten dituzten seinaleak markatu ditugu testuetan, eta denbora-informazioaren tratamendu automatikoan erabilgarri zaigun informazio linguistikoa kodetzeko balioak definitu ditugu. Era berean, gidalerro horien egokitasuna neurtu dugu etiketatzaileen lana ebaluatuz.

3. EusTimeBank corpuseko testuak etiketatu ditugu EusTimeML baliatuta. EusTimeBank corpusak 164 dokumentu ditu: euskarazko albisteak eta historia-narrazioak. Dokumentuak eskuz etiketatu ditugu eta horietako 60 euskarazko denbora-informazioa prozesatzeko tresnak (EusHeidelTime, bTime eta KroniXa) entrenatzeko eta ebaluatzeko erabili ditugu.

4. Denbora-adierazpenak automatikoki identifikatzeko, sailkatzeko eta normalizatzeko, EusHeidelTime tresna sortu dugu. EusHeidelTime erregeletan oinarritzen da eta denbora-adierazpenak identifikatzeaz gain, horiek kronologiako zein punturi egiten dioten erreferentzia adierazten du.

5. KroniXa tresnak euskarazko testuetatik denbora-lerroak sortzen ditu testuko gertaerak gertatzen diren uneen arabera antolatuta. Horretarako, EusHeidelTimek identifikatzen duen informazioa bTimek erauzten duenarekin (gertaeren eta denbora-erlazioen identifikazioa eta sailkapena) eta dependentzia sintaktikoetatik lortutakoarekin uztartuta denbora-lerroak sortzeko beharrezko informazioa lortu dugu.

Hala, urrats horiei jarraituta, euskarazko testuetako denbora-informazioa identifikatu, aztertu eta, hori baliatuta, denbora-lerroak sortzeko prozesua bete dugu.