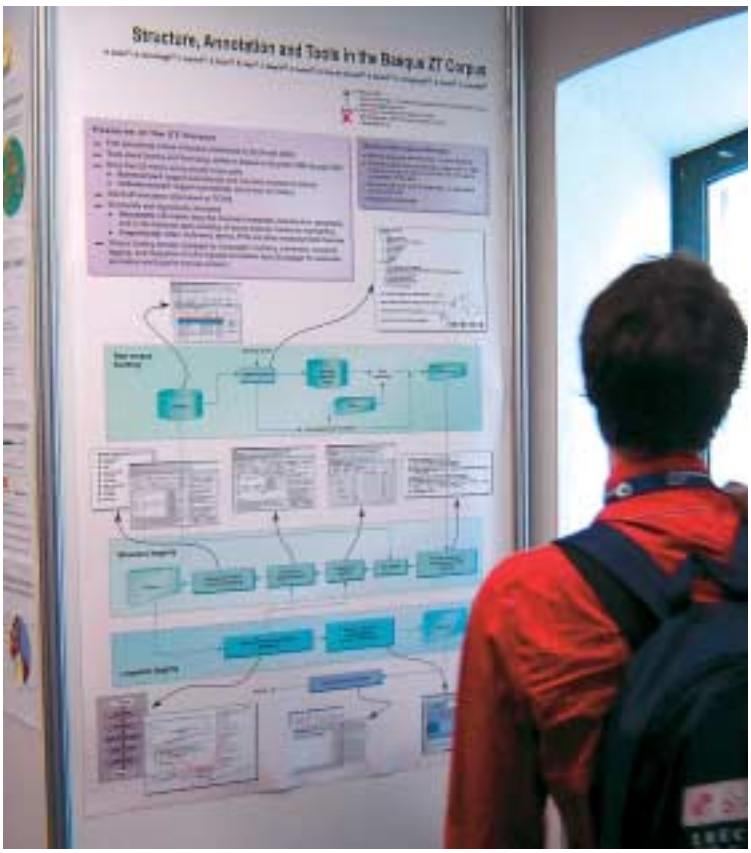
Zientzia eta Teknologiaren Corpusa-ren Interneteko lehen bertsioa

Gurrutxaga Hernaiz, Antton
Elhuyar Hizkuntza Zerbitzuak

Abenduaren 14an, EHUko IXA Taldeak eta Elhuyar Fundazioak *Zientzia eta Teknologiaren Corpusa*-ren Interneteko bertsioa aurkeztuko dute. Euskarazko lehen corpus berezi edo espezializatua da. Zientzia eta teknologiaren alorreko euskarazko testuen bilduma egituratu eta etiketatua da, eta alor horietako euskararen erabilera ikertzeko baliabidea izatea du helburu nagusia.



A. GURRUTXAGA

Zientzia eta Teknologiaren Corpus-aren aurkezpena hizkuntza-baliabideei buruzko LREC biltzarrean (Genoa, 2006).

CORPUSA ELIKATZEKO, 1990-2002 BITARTEAN argitaratutako zientzia eta teknologiaren alorreko obrak hartu dira kontuan. Corpusa eremuaren (jakintza-alorraren) eta generoaren arabera sailkatuta dago.

Corpusa etiketatuta dago, bai testuaren egiturari eta formatuari dagokionez, bai linguistikoki. Etiketatze linguistikoa egiteko, euskara automatikoki prozesatzeko teknologia aurreratua erabili da (IXA taldearen Eustagger etiketatzailea). Testuko hitz bakoitzaren lema eta kategoria/azpikategoria etiketatuta dira. Corpusaren bertsio honetan, 8 milioi hitz daude, eta horietatik, 1,6 milioi hitz eskuz berrikusi, desanbiguatu eta zuzendu dira. Corpusa XMLn etiketatuta dago, eta TEI estandarrari jarraitu diogu.

Corpusa kontsultatzeko interfaze ahaltsua antolatu dugu, eta erabiltzaileak era askotako bilaketa bakunak eta konplexuak egiteko aukera izango du, horretarako parametro-multzo zabala erabiliz: lema, testu-forma, kategoria, eremua, generoa, corpus-atala (eskuz zuzendua/corpus osoa...). Emaitzak bi eratakoak izan daitezke. Batetik, bilagaiaren testuinguru laburrak (KWIC) eta testuinguru hedatuak; eta, bestetik, informazio kuantitatiboa, taula eta grafikoetan emana (maiztasunak, agerkidetzak, eremu edo generoaren araberako banaketa, eta abar).

Corpusa **www.ztcorpusa.net** gunean egongo da kontsultagai. Horrez gain, 2007tik aurrera ELDAren baliabideen artean egongo da, ustiapen komertzialerako eskuragarri, lizentzia bidez.

Corpusaren lehen bertsio honetan sartu diren testuak formatu digitalean jaso ditugu hainbat hornitzailerengandik, haiekin sinatutako hitzarmenei esker. Bihoazkie denei ere gure esker beroenak.

Zientzia eta Teknologiaren Corpusa proiektua Hizking21 ikerketa estrategikoko proiektuaren barnean hasi zen egiten. Hizking21 proiektuak honako laguntza hauek jaso ditu: Eusko Jaurlaritzaren Industria Sailaren Etortek programa (2002-2004) eta Gipuzkoako Foru Aldundiaren Gipuzkoako Zientzia, Teknologia eta Berrikuntza Sarea programa (2004). Bestetik, *Zientzia eta Teknologiaren Corpusa*-k Eusko Jaurlaritzaren Kultura Sailaren 2005eko Euskara eta Teknologia Berriak programaren laguntza ere jaso du.