Koherentziazko erlazioak: marko teorikoa eta corpusaren deskribapena

Mikel Iruskieta, Arantza Diaz de Ilarraza & Mikel Lersundi
Ixa taldea, UPV/EHU

Laburpena

Lan honetan Hizkuntzalaritza Konputazionalaren diziplinan eta erlaziozko diskurtso-egituraren aztergaian koherentzia aztertzeko gehien erabiltzen den Egitura Erretorikoaren Teoria (Rhetorical Structure Theory edo RST) aurkeztuko dugu eta baita teoria horri esker euskaraz deskribatu den Euskal RST Treebanka ere. Corpusa lau hizkuntzalarik etiketatu dute eta etiketatzaileen artean sorturiko ezadostasunak ebatzi ditu epaile batek. Ondorioz, corpus horretan zientzia-artikuluetako 60 testu-laburpenen ereduzko erlazio-egitura deskribatu dugu, baita corpusaren ezaugarriak, etiketatzeko irizpideak eta etiketatzaileen adostasuna ere. Bestalde, corpus horretan egin daitezkeen kontsultak zeintzuk diren aipatzen ditugu. Nabarmentzekoa da corpus honetan egin daitezkeen kontsultak RSTn eskaintzen diren beste corpus batzuk baino aurreratuagoak direla eta kontsulta horiek garatzeko erabili ditugun programak beste hizkuntzetarako ere balio dutela.

0. **Sarrera***

Diskurtsoaren egitura Hizkuntzalaritza Konputazionalean deskribatzean, bi fenomeno lantzen dira batez ere: a) erreferentziazkoa: korreferentzia ebazpena (Goenaga et al 2012; Recasens et al 2010; Mitkov 2002) eta b) erlaziozkoa: koherentzia erlazioen esleipena (Iruskieta 2014; Prasad et al 2007; Asher & Lascarides 2003; Mann & Thompson 1988). Lan honetan b) fenomenoaz arituko gara eta zientziako testu errealen koherentzia-erlazioak izango ditugu ikergai.

Lan honetako helburua da Hizkuntzalaritza Konputazionalean erlaziozko diskurtso-egitura aztertzeko baliagarria den teorietako bat, Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory* edo RST), aurkeztea. Baita teoria

* Lan hau burutu dugu proiektu hauen laguntzei esker: NewsReader: ICT Call 8 FP7-ICT-2011-8-316404 (Europako Batzordea). Ber2tek: IE12-333 (Eusko Jaurlaritza).

horri esker euskaraz deskribatu den Euskal RST *Treebank*a deskribatzea eta bertan egin daitezkeen kontsultak zeintzuk diren zehaztea ere.

Erlaziozko diskurtso-egitura edo erlazio-egitura[1] esaten diogu testuan koherentzia-erlazio guztiek osatzen duten egiturari. Euskaraz hizkuntzaren ikuspegi deskriptibo gramatikaletik edota didaktikotik koherentzia-erlazioen barnean dauden fenomeno nabarmenenak (diskurtso-markatzailez seinalaturiko erlazio esplizituak eta erlazio semantikoak) aztertu dira batez ere (Alberdi & Garcia 2012; Esnal 2008; Euskaltzaindia 2005, 1999, 1994, 1990; Larringan 1995, besteak beste). Ikuspegi horietan, erlazio-egiturako fenomeno nabarmen horien deskripzioa nahikoa izan daiteke. Hizkuntzaren Prozesamenduan, ordea, konputagailuaren bitartez erlazio-egitura ezagutzeko koherentzian parte hartzen duten fenomeno guztiak (batez ere erlazio inplizituak[2] eta pragmatikoak) zehaztasunarekin deskribatu behar dira automatikoki errekonozitu eta tratatu ahal izateko. Aipatu ditugun fenomeno guztiak aurkezteko van Dijken (1980) laneko zenbait adibide erabiliko ditugu.

(1) Sarrera erosi dut eta nire aulkira joan naiz.
(2) #Peter zinemara joan zen. Berak begi urdinak ditu.
(3) John gaixorik dago. Gripea dauka.
(4) Johnek ezin du etorri. Gaixorik dago.

(1) adibideko koherentzia azaltzeko —bi esaldien artean dagoen erlazio-egitura ulertzeko edo esaldiko gertaeren testuingurua zein den jakiteko— egin behar dugun ahalegin berezia koherentziaren bi mailak (lokalean eta globalean) eta bi maila horien arteko harremana zein den esan behar da: izan ere, koherentzia lokala (esaldien artean dagoen erlazioa: tiketa erostearen eta aulkira joatearen artean) ulertu ahal izateko jakin behar dugu bi esaldi horien arteko erlazioa makroegiturak edo gaiak (zinemara joatea) ezartzen duela. Horrela, perpausen artean *eta*k juntatutako gertaera-sekuentzia (SEKUENTZIA erlazioa) dagoela ulertuko dugu. (2) adibideko esaldiak, ordea, ez dira koherenteak maila lokalean, nahiz eta pertsona berari buruz diharduten. Pentsa lezake norbaitek koherentzia falta, kasu honetan, erlazio inplizitua delako gertatzen dela; baina koherentzia eza irakurleak makroegituraren eta mikroegituraren arteko harremana ez bilatzean datza. Bestalde, (3) adibideko bigarren esaldiak lehenengoa zehazten duenez, nahikoa da adieraziriko bi egoeren arteko (*gaixorik egon* eta *gripe izan*) erlazioa (ELABORAZIOA erlazioa) zein den azaltzea; horrelako kasuetan erlazio semantiko inplizitua dagoela esaten da, ez baitago diskurtso-markatzailerik. Azkeneko adibidean (4), aldiz, koherentziazko erlazio bat baino

[1] Beste ikuspegi batetik Sainzek (2001: 158) *erlazio-egitura* esateko *proposizio-egitura* terminoa erabili izan du, guk *erlazio-egitura* darabilgu termino zabalagoa delako eta darabilgun marko teorikoan egitura erlazioen bidez deskribatzen delako.

[2] Erlazio inplizituei bibliografian modu askotan esaten zaie: *implicit*, *unsignal*, *unmarked* edo *underspecific*.

gehiago egon daitezke: i) adieraziriko bi egoeren artean (*ezin etorria* eta *gaixorik egotea*) erlazio semantikoa balego, pentsatu beharko genuke gaixotasuna dela etorri ezinaren arrazoia edo kausa (KAUSA erlazioa); ii) baina erlazio hori ez da bi esaldi horien arteko koherentzia-erlazio bakarra, ezta erlaziorik nabariena ere; izan ere, badakigu gure gizartean gaixorik egotea (gripea izatea) bilera batera edo lanera ez etortzeko justifikazio nahikoa dela (JUSTIFIKAZIOA erlazioa). Interpretazio hori egokiena denean, bi esaldi horien arteko koherentzia-erlazioak entzulearengan efektu bat eragiteko egiten da, ez baitago semantikoa den erlaziorik bi esaldion artean, eta horregatik esaten da erlazio pragmatikoa edo erretorikoa dagoela bi esaldi horien artean.

Aipatutako fenomeno berezi horiek (erlazio anbiguoak eta pragmatikoak) oso arruntak dira testu errealetan eta, horregatik, Hizkuntzalaritza Konputazionaleko hainbat atazatan testu errealak eta osoak aztertzeko erlazio semantikoez gain, erlazio pragmatikoak edota erlazio inplizituak deskribatzeko marko teorikoa behar da. Hori dela eta, lan honetan aztertuko dugun Egitura Erretorikoaren Teoriak (*Rhetorical Structure Theory*, aurrerantzean, RST) testuaren egitura edo koherentzia deskribatzen du (Mann & Thompson 1987), aipatutako fenomeno berezi horiek ere kontuan hartuz. Marcuren (2000) arabera, RST da hizkuntzaren sorkuntza automatikoan (*text generation*) dabiltzanen artean teoriarik erabiliena. Testuak sortzeko erabiltzeaz gain, testuak prozesatzeko erlazio-egitura deskribatzeko balio duen teoria aplikagarria asmatu zuten, baita beste aplikazio askotarako ere (Taboada & Mann, 2006).

RSTren bidez koherenteak diren testu mota gehientsuenak (Marcu 2000) deskriba daitezke, hizkuntza batean edo bestean izan (Mann & Thompson 1987). Horren erakusgarri da, besteak beste, RST erabiliz deskribatu diren alor eta hizkuntza ezberdinetako testu-multzoak: i) ingelesez kazetaritza-testuak (Carlson et al 2002) eta iragarkiak, gutunak, aldizkarietako artikuluak, artikulu zientifikoak, film- eta liburu-kritikak, iritzi-artikuluak eta kazetaritza-testuak (Taboada & Renkema 2011); ii) alemanez kazetaritza-testuak (Stede 2004); iii) portugesez informatikari buruzko testu zientifikoak (Pardo & Seno 2005) eta irakasleen ahozko testuak (Antonio 2004); eta iv) gaztelaniaz gai batzuetako zientzia-testuak: Astrofisika, Sismologia, Ekonomia, Legedia, Hizkuntzalaritza, Matematika, Medikuntza, Psikologia eta Sexualitatea (da Cunha et al 2011).

Testuaren egitura deskribatzeko, zehaztu behar da testuaren osagaiak zein diren eta osagaien artean zein harreman edo erlazio dagoen. Testuak aztertzerakoan, testua zatitzea da abiapuntua. Testua sortzean, aldiz, zatiak elkarrekin erlazionatzea izango da abiapuntua. Oinarrizko lan honetan helburua testua aztertzea denez, lehenengo urratsa testua zatitzea edo segmentatzea izango da.

Testua segmentatu ondoren, segmentuek elkarren artean dituzten koherentziazko erlazioak ezartzen dira. Unitate bat beste segmentu bati lotzeko, unitate horiek koherentetzat hartu behar ditu irakurleak; izan ere, segmentuen artean koherentziarik ez badago, ez da segmentuen arteko koheren-

tzia-erlaziorik egongo (*non-sequitur*) edo, zehazkiago esanda, irakurleak ez du ulertuko zein den segmentuen arteko koherentzia-erlazioa, (2) adibidean gertatzen den bezala. Zenbait erlaziok irakurlearengan duen eraginagatik esaten zaie koherentzia-erlazioei erlazio erretoriko (*rhetorical relation*, Pardo 2005). *Erretoriko* hitza, ordea, zentzu zabalean ulertu behar da Ghorbel et al.en (2001) arabera; izan ere, erlazio semantiko, pragmatiko, logiko edo domeinuaren araberako erlazioek osatzen dituzte aipatutako erlazio erretorikoak. Beraz, RSTrekin testu koherenteak aztertzean ez da *non-sequitur* edo hutsunerik egongo eta testuko segmentu guztiek loturaren bat izan behar dute beste segmenturen batekin. Lotura hori zein den adierazteko erlazio erretorikoez lotzen dira segmentuak. RSTn testu koherenteko erlazio-egiturako segmentu edo diskurtso-unitate guztiak lotuta egongo dira zuhaitz-egituran.

Hierarkiari dagokionez, erlazio-egitura mota bi daude: i) garrantzia bereko segmentuez osaturiko erlazio-egiturak (Prasad et al 2007) eta ii) garrantzi ezberdinez osaturiko erlazio-egitura hierarkikoak (Mann & Thompson 1987 eta Lascarides & Asher 2007). Idazleak testuaren mezua edota intentzioa komunikatzeko erabiltzen dituen segmentu batzuk gako izaten dira bere ideia nagusiak azaltzeko, eta beste unitate batzuk, bigarren mailakoak, ideia nagusien xehetasunak azaltzeko. Idazlearen ideia nagusiak RSTn garrantzi ezberdina duten segmentuen bidez deskribatzen da. Hortaz, RSTk ezberdindu egiten du diskurtso-unitate batzuk beste batzuk baino garrantzitsuagoak direla diskurtso-egituran. Horiei nukleo-unitate (*nuclear unit*, N) esaten zaie eta besteei satelite-unitate (*satellite unit*, S). Unitate bat nukleo ala satelite den esateari nukleartasuna (*nuclearity*) ebaztea deritzo. Nukleartasuna koherentziarekin lotuta dago, testuko nukleo-unitateak bakarrik irakurrita ulertzen baita idazleak testuarekin komunikatu nahi duena. Nukleoz osaturiko testuak kohesioan edota perpaus-joskeran (*clause combining*) hutsuneak izan baditzake ere, koherentea izango da. Alderantziz, ordea, ez; testuko nukleo-unitateak kenduta nukleorik gabeko testu-hondarra irakurriko bagenu, ulergaitza eta ezkoherentea litzateke (Mann & Thompson 1988). Horregatik da diskurtsoaren nukleartasuna oso baliagarria laburpen automatikoko atazetan (Bosma 2008). Horrela, nukleartasunak zuhaitz itxurako egiturari hierarkia ezartzen dio, erlaziozko diskurtso-egitura deskribatzean. Nukleo bakarreko erlazioetan, (3) eta (4) adibideak, satelitea da nukleoari erlazio erretorikoz lotzen zaiona: (3) adibidean ELABORAZIOA denez, zehatza (*gripea izatea*) lotzen zaio orokorrari (*gaixorik egoteari*) eta (4) adibidean JUSTIFIKAZIOA denez, arrazoia da (*gaixorik egotea*) lotzen zaiona nukleo-unitateko (*ezin etortzeko*) baieztapenari. Erlazio horiei, nukleo-satelitez (N-S) loturikoei, *erlazio hipotaktiko* deritze eta, bestalde, nukleo-nukleoz (N-N) osaturiko nukleo anitzeko erlazioei (*multinuclear*) *erlazio parataktiko*, (1) adibidean SEKUENTZIA erlazioa, adibiderako.

Nukleartasuna eta erlazio mota kontuan hartuz RSTn 30 erlazioko zerrenda proposatu da. Zerrenda horri «RSTko sailkapen klasiko heda-

tua»[3] esaten zaio (Mann & Taboada 2010). Hauek dira sailkapen horretako erlazio guztiak:

i) Erlazio nukleoaniztunak (N-N) nukleoz bakarrik osatuta daudenez, erlazio nukleobakarretan ez bezala, nukleo-unitate guztiek garrantzi bera dute eta, beraz, hierarkiarik ez dago lotzen dituzten unitateen artean. Hauek dira erlazio nukleoaniztunak: LISTA, DISJUNTZIOA, BATERATZEA, BIRFORMULAZIOA-NN, SEKUENTZIA, KONTRASTEA eta KONJUNTZIOA.

ii) Erlazio nukleobakardunek (N-S) garrantzi ezberdineko proposizioak lotzen dituzte. Nukleo-unitateak predikazioaren edo komunikatu nahi denaren zatirik garrantzitsuenak dira; satelite-unitateek, berriz, gaia (zeina nukleo-unitatean adierazten den) hobeto edo zehatzago ulertzen laguntzen dute. Erlazio hauek lotzen dituzten segmentuei dagokienez, bai esaldi barruan, bai esaldien edo paragrafoen barruan egon daitezke eta bi motatakoak dira:

   a) Edukizkoak. Edukizko erlazioetako (*subject matter*) unitateen loturak izaera semantikoa du; hau da, idazlearen intentzioa da irakurleari jakinaraztea erlazio-unitateen artean zein erlazio dagoen: ELABORAZIOA, METODOA, ZIRKUNSTANTZIA, ARAZO-SOLUZIOA, BALDINTZA, AUKERA, ALDERANTZIZKO BALDINTZA, EZ-BALDINTZAILEA, INTERPRETAZIOA, EBALUAZIOA, ONDORIOA, KAUSA eta HELBURUA.
   b) Aurkezpenezkoak. Aurkezpenezko erlazioetako (*presentational*) unitateen loturek, ordea, izaera erretorikoa dute; beraz, unitateen arteko loturek irakurlearengan efektu jakina eragiteko asmoa dute: PRESTATZEA, TESTUINGURUA, AHALBIDERATZEA, MOTIBAZIOA, EBIDENTZIA, JUSTIFIKAZIOA, ANTITESIA, KONTZESIOA, BIRFORMULAZIOA eta LABURPENA.

RSTn erlazio erretorikoak definitzeko baldintza hauek hartzen dira kontuan:

i) Erlazioko unitate bakoitzak bete behar dituen baldintzak, dela nukleoa, dela satelitea.
ii) Erlazioko unitate guztiek (nukleoak eta sateliteak batera) bete behar dituzten baldintzak.

---

[3] Erlazio erretorikoen sailkapenari dagokionez, sailkapen ezberdinak daude eta, marko teorikoan erlazio berriak proposatzeko murriztapenik ez badago ere, guk RSTko sailkapen hedatua aukeratu dugu. Sailkapen horren euskarazko definizioak eta adibideak RSTko webgunean kontsulta daitezke: [http://www.sfu.ca/rst/07basque/definitions_RST_Basque.html]. Sailkapen hedatua hizkuntza ezberdinetako testuak etiketatzeko erabili izan da. Adibidez, O'Donnellek (2000) testuak RSTrekin aztertzeko egindako tresnak (RSTTool) ingeleserako, frantseserako eta gaztelaniarako sailkapen hedatuak dakartza.

iii) Idazleak irakurlearengan eragin nahi duen efektuari buruzko baldintza.[4]

Baldintza horiek zertan oinarritzen diren erakusteko, aipatu ditugun adibideetako erlazioak RSTn nola definitzen dituzten erakutsiko dugu 1. taulan.[5]

1. TAULA

RSTko erlazioen definizioak

Sekuentzia	Arauak *N* bakoitzeko		Efektua
	N guztien artean segida erlazioa dago		*N* guztien arteko segida erlazioa ezagutzen du irakurleak
Elaborazioa	Arauak *Sn* eta *Nn*	Arauak *S*-*N*n	Efektua
	Ez dago baldintzarik	*N*n aurkeztutako gaiaren edo egoeraren ezaugarriren bat garatzen da *S*n edo *N*tiko inferentzia aurkezten da *S*n, erlazio hauen arabera: — multzoa :: kidea — abstraktua :: adibidea — osoa :: zatia — prozesua :: urratsa — objektua :: atributua — orokorra :: espezifikoa	*S*n aurkeztutako egoerak *N*ko ezaugarriren bat garatzen duela onartzen du irakurleak. Irakurleak garatutako elementua edo gaia identifikatzen du
Justifikazioa	Ez dago baldintzarik	Irakurleak *S* ulertzean, idazleak *N* aurkezteko egokitasuna areagotzen da	Irakurleak idazleari *N* aurkezteko egokitasuna onartzen dio

[4] Unitateen artean dagoen koherentzia-erlazioaren baldintzak erakusten digu zein den idazlearen intentzioa. Erlazio bakoitzaren efektua irakurtzean, formula hau izan behar da gogoan: «*It is plausible to the analyst that it was plausible to the author that*» (Mann & Taboada 2010).

[5] SEKUENTZIA erlazio nukleoaniztuna definitzean, nukleoetako egoeren artean segida erlazioa dago. (1) adibidean zineman film bat ikusteko, lehendabizi *sarrera erosi* behar da eta ondoren *aulkira joan* behar da. Irakurleak efektu hori, egoeren arteko erlazioak sorrarazten diona, ezagutzen du.

ELABORAZIOA erlazioa definitzean, nukleoaren eta satelitearen arteko erlazioan sateliteko egoera nukleotik inferi daiteke eta irakurleak bi egoeren arteko efektua identifikatzen du —(3) adibidean nukleotik, *gaixorik egon* egoera orokorretik sateliteko egoera inferi daiteke, gripea gaixotasun mota bat delako—.

JUSTIFIKAZIOA erlazioa definitzean, nukleoaren eta satelitearen arteko erlazioan baldintza da idazleak nukleo-unitatean aurkezturiko egoeraren —(4) adibidean *lanera ez etortzea*ren— azalpena ematen duela eta satelite-unitateko —*gaixorik egotea* (gripeaz gaixorik dagoela)— azalpen horretan nukleo-unitateko egoeraren zergatiak aipatzen direla. Bestalde, unitate bi horien erlazioak irakurlearengan duen efektua da irakurleak idazleari nukleo-unitateko egoera aurkezteko zilegitasuna onartzen diona.

Horiek horrela, azpimarratu nahi dugu koherentzia RSTn erlazio erretorikoekin soilik deskribatzen dela. Ez da bestelako kontzepturik behar, ezta diskurtso-markatzaileena ere; bada, erlazio erretorikoa —(4) eta (3) adibideetan— diskurtso-markatzailerik gabe ere interpretatu behar du irakurleak erlazio inplizituen kasuan (Oates 1999).

Eskuzko etiketatzeaz gain, testua automatikoki sortu edota prozesatu nahi bada, konputagailuan testu mailako ezaugarriak zeintzuk diren zehaztu behar dira. Ondo egituratutako testuen ezaugarri nagusia koherentzia da; hortaz, konputagailuak koherentzia prozesatzeko informazioa behar du. Testuak aztertzen dituzten konputagailuei informazio egokia eman behar zaie koherentzia egokiro deskribatzeko eta, horrela, konputagailuak testuaren erlazio-egitura prozesatu edota sortu ahal izango du. Ildo horretan aipagarriak dira diskurtsoa era automatikoan aztertzeko diseinatu eta inplementatu diren programak edo diskurtso-analizatzaile automatikoak, besteak beste, i) japonierarako (Sumita et al 1992); ii) ingeleserako (Corston-Oliver 1998; Marcu 2000; Hanneforth et al 2003; Mahmud & Ramsay 2005) eta iii) Brasilgo portugeserako (Pardo et al 2004).

1. **Metodologia**

Erlaziozko diskurtso-egitura deskribatzeko marko teorikoa aurkeztu ondoren, RSTrekin euskarazko corpusa nola deskribatu dugun azalduko dugu atal honetan. Horretarako, corpusa (1.1. azpiatala), etiketatzaileen deskribapena (1.2. azpiatala), etiketatze faseak (1.3. azpiatala) eta etiketatzearen ebaluazioa (1.4. azpiatala) modu laburrean azalduko ditugu.

1.1. *Corpusa*

Euskarazko lehen RSTko corpusa eratzeko, hautatu ditugun irizpideak honako hauek dira:

i) Erlazio erretorikoen azterketan domeinuen eragin orekatua izateko erabaki dugu domeinu aniztuna izatea eta domeinu bakoitzetik testu kopuru bera hartzea. Lan honetarako erabili dugun hiru domeinu ezberdineko corpusa 2. taulan ageri den bezala osatuta dago: a) Medikuntzako testuak *Gaceta Médica de Bilbao* aldizkariko artikuluetako laburpenak dira. b) Terminologiako testuak HAEE-IVAPek eta UZEIk antolatuta, Donostian 1997 urtean egin zen Nazioarteko Terminologia Biltzarreko aktetako hitzaldi eta komunikazioen laburpenak dira. c) Zientziako testuak Euskal Herriko Unibertsitateko (UPV/EHU) Zientzia eta Teknologia Fakultateak

2008ko maiatzean antolaturiko Zientzia eta Teknologia Fakultateko 1. Ikerkuntza Jardunaldiak izeneko komunikazioetako laburpenekin osatu dugu.

2. TAULA

Aztergai dugun euskarazko corpusaren deskribapena

Domeinua	Azpicorpusa	Testuak	Esaldiak	Hitzak
Medikuntza	GMB	20	198	3.010
Terminologia	TERM	20	253	5.664
Zientzia	ZTF	20	352	6.892
Guztira		60	803	15.566

ii) Testuak ondo egituratuta egotea eta laburrak izatea erabaki dugu. Ondo egituratutako testuak interesatzen zaizkigu, batetik, aukeratutako testu-moten erlazio-egitura zein den jakiteko eta, bestetik, etiketatzaileen arteko adostasuna ahalik eta handiena izateko. Halaber, testu laburrak ere interesatzen zaizkigu, erlaziozko diskurtso-egiturak eskuz konpara eta oso zehatz ebalua daitezkeelako. Horretarako, artikulu zientifikoetako laburpen-testuak dira egokienak.

iii) Testuak hizkuntza batean baino gehiagotan egotea erabaki dugu, azterketa kontrastiboak egin ahal izateko eta itzulpen automatikoko atazetan erabili ahal izateko (Iruskieta et al 2014a; Iruskieta & da Cunha 2010a, 2010b; da Cunha & Iruskieta 2010).

1.2. *Etiketatzaileen eta epailearen deskribapena*

Corpuseko lau etiketatzaileak hizkuntzalariak izan dira. Gehienek hizkuntzako beste maila batzuk (morfosintaxia, sintaxia eta semantika) etiketatu dituzte; ez, ordea, diskurtso mailako fenomenorik. Beraz, erlazio-egitura etiketatzean izan dute RSTren berri. RST aurkeztu ostean, etiketatzeko zenbait irizpide azaldu eta *RSTTool* aurkeztu zaie. Noizean behin egitura batzuk irudikatzeko zalantzak argitu izan badira ere, ez da entrenamendu-faserik egon.

Epaileari dagokionez, RSTrekin testu gehien etiketatu eta ebaluatu dituen etiketatzailea izan da. Epaileak beste etiketatzaileen anotazioak ikusi aurretik etiketatu du fase bakoitza eta behin etiketatzaile guztien anotazioak bilduta, bere eta etiketatzaileen artean ezadostasuna egon bada, irizpideak eraiki ditu eta irizpide horiei jarraituz, ezadostasun kasuen eta irizpideak jarraitzen ez dituzten kasuen harmonizazio-prozesua gauzatu du (Iruskieta 2014).

1.3. *Corpusaren etiketatze-faseak*

Testua erlazio-egituraz eta erlazio-seinaleekin aberasteko prozesua lau fasetan egin dugu. Fase bakoitza ebaluatu eta epaile batek aurretik zehaztu-riko erabakiei jarraituz urre-patroia sortu du, etiketatzaileak fase bakoitzean oinarri beretik hasteko eta fase bakoitzeko arazoak deskribatzeko. Faseak lau hauek izan dira:

i) Segmentazioa: segmentatzeko zenbait irizpidetan oinarrituz etiketatzaileei testua oinarrizko diskurtso-unitateetan (EDU) zatitzeko eskatu zaie. Orokorrean EDU bat aditza duen perpaus adjuntua edo perpaus independentea da (Iruskieta et al 2011a, 2011b).
ii) Makroegituraren identifikazioa: etiketatzaileek testuko unitaterik garrantzitsuena edo unitate zentrala (UZ) aukeratu dute erlazio erretorikoak aukeratu aurretik (Iruskieta 2014); izan ere, makroegiturak koherentzia lokalean eragiten du (Iruskieta et al 2014b; van Dijk 1980).
iii) Erlazio-egituraren errepresentazioa: unitate zentrala zein den kontuan izanik, EDUen eta unitate-multzoen arteko erlazioak zehazten dira. Egitura erretorikoa eraikitzeko unitateak era inkrementalean eta moduluka erlazionatuko dira Pardoren (2005) lanean proposatu den bezala: lehendabizi esaldi barneko unitateak erlazionatuz, ondoren paragrafo barrukoak eta, bukatzeko, testu barrukoak (Iruskieta et al 2013b).
iv) Erlazio-seinaleak etiketatzea: etiketatzaileetako batek erlazio erretorikoak seinalatzen dituzten elementuak etiketatu ditu (Iruskieta 2014), Taboada & Das (2013) lanean proposatutakoari jarraituz.

1.4. *Etiketatzearen ebaluazio-emaitzak*

Corpus honetako etiketatze-fase ezberdinetan fidagarritasunaren emaitzak zehatz azaldu ditugu Iruskietaren (2014) tesi-lanean. 3. taulan corpus osoa etiketatu duten bi etiketatzaileen arteko adostasuna laburbilduko dugu.

Zerbait nabarmentzekotan, esan dezakegu etiketatutakoaren fidagarritasuna asko jaisten dela erlazio erretorikoetan; izan ere, adostasuna beste fenomenoetan dagoen adostasuna baino nabarmen txikiagoa da. Baina hori ez da ez euskararen diskurtso-egituraren arazoa, ez etiketatze-egitasmo honetakoa; izan ere, antzeko ezaugarriak dituzten etiketatze-egitasmoetan (ingelesez: Carlson et al 2001; nederlanderaz: van der Vliet 2011) lortu diren emaitzak lortu dira.

3. TAULA

Etiketatze maila guztietako adostasuna

	Testu segmentazioa	Ideia nagusia	Erlazio erretorikoak	Kausazko seinaleak
F-neurria	% 81,35	% 81,67	% 61,47	% 76,82

2. **Euskal RST *TreeBank*a**

Etiketatze-fase horiek jarraituz sortu dugu hurrengo ataletan deskribatuko dugun Euskal RST *TreeBank*a.[6] Webgunean sartzean, corpuseko erlazio-egitura fenomeno ezberdinetatik abiatuz kontsultak fitxategiz fitxategi egin daitezke: esaterako, fitxategi horren erlazio-egituraren irudia ikus edo lor daiteke *jpg* formatuan, erlazio-egitura *xml* eta *rs3* formatuan, testu-fitxategiak *txt* formatuan, erlazio-egituraren seinaleak *rhetdb* formatuan eta informazio morfosintaktikoz aberastutako dokumentuak *kaf* formatuan. Azpimarratu nahi dugu euskarazko *TreeBank*ean RSTko bibliografiako beste lanetan baino kontsulta-zerbitzu osoagoa eskaintzen dugula; gainera, beste hizkuntzetarako (ingelesa, gaztelania eta portugesa)[7] ere baliagarri dela frogatu dugu.

2.1. *Segmentuak: oinarrizko diskurtso-unitateak (EDU) eta ideia nagusiak (UZ)*

Atal honetan *(http://ixa2.si.ehu.es/diskurtsoa/segmentuak.php)* corpuseko fitxategi guztien izenen gainean klik egiten bada, testu horretako segmentazioa edo oinarrizko diskurtso-unitateak (EDUak) kontsulta daitezke, baita testuko ideia nagusia edo unitate zentrala (UZ) ere; adibidez, 4. taulan GMB0301 testuko segmentazioa erakusten dugu. Testuak zazpi EDU ditu eta azkena, EDU_7a, UZtzat hartu du etiketatzaileak; beraz, etiketatzailearentzat EDU hori da testu horretako segmenturik garrantzitsuena. Webgunean EDU horri sakatuz gero, GMB0301 testuko unitate zentralari lotutako erlazio erretorikoak agertuko dira.

[6] http://ixa2.si.ehu.es/diskurtsoa/

[7] Ingelesez eta gaztelaniaz: [http://ixa2.si.ehu.es/rst/] eta portuguesez: [http://ixa2.si.ehu.es/pt/]

4. TAULA

Diskurtso-unitateen kontsulta, GMB0301

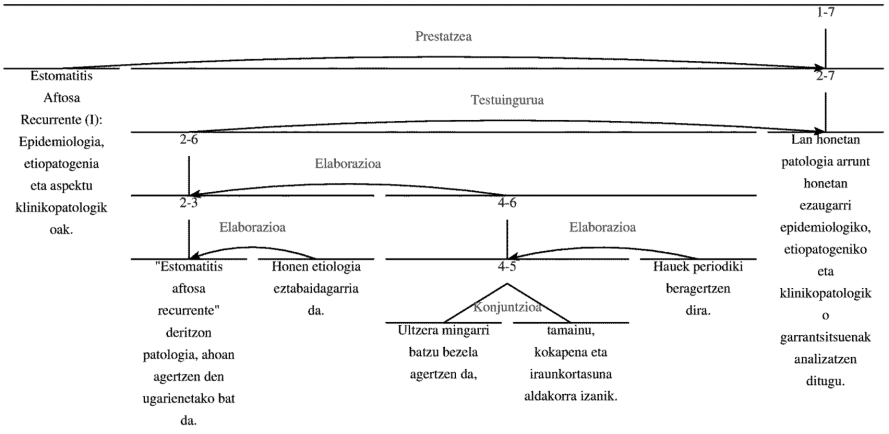
EDU	GMB0301-GS.rs3 (7) *Segmentuak*
1	Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.
2	«Estomatitis aftosa recurrente» deritzon patologia, ahoan agertzen den ugarienetako bat da.
3	tamainu, kokapena eta iraunkortasuna aldakorra izanik.
4	Honen etiologia eztabaidagarria da.
5	Ultzera mingarri batzu bezela agertzen da.
6	Hauek periodiki beragertzen dira.
7 (UZ)	*Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.*

2.2. *Erlazio erretorikoak*

Beste atal honetan bi motatako kontsultak egin daitezke: i) dokumentu baten gainean (http://ixa2.si.ehu.es/diskurtsoa/fitxategiak.php) klik egin eta bere diskurtso egitura irudikatu eta ii) testu baten erlaziozko diskurtso-egituran erlazio guztiak bilatu (http://ixa2.si.ehu.es/diskurtsoa/zuhaitzak.php). Bilaketa zabal daiteke corpus osora edo azpicorpus batera (GMB, edo TERM, edo ZTF). Esaterako, 1. irudian GMB0301 testuaren erlazio-egituraren errepresentazioa erakusten dugu.

Bestalde, corpuseko erlazioen kontsultari dagokionez, KAUSA erlazioaren lehenengo lau agerpenak erakusten ditugu 5. taulan. Taula horretan lehenengo hiru zutabeetan diskurtso-unitateen ordena[8] eta noranzkoa deskribatzen dira. Segmentuen ordena (ezkerrekoa eta eskumakoa) testuko ordenarekin bat datorrenez, erlazioaren nukleartasuna bigarren zutabean definitzen da: erlazioak NS (ezkerrean nukleoa eskuman satelitea) ordena badu, geziak ezkerrerako noranzkoa (←) izango du, gezia nukleoari zuzenduz; SN ordena badu, geziaren noranzkoa eskumarakoa (→) izango da. Laugarren eta bosgarren zutabeetan erlazioa eta adibidearen iturria zehazten dira.

[8] Erlazio-seinaleak diskurtso-unitateetan nabarmenduta daude beltzez eta azpimarratuta.



1. IRUDIA

Erlaziozko diskurtso-egituraren kontsulta, GMB0301

5. TAULA

Euskal RST TreeBankeko KAUSA erlazioaren kontsulta adibidea

Erlazioa: kausa (27)				
Ezkerrekoa	NS	Eskumakoa	Erlazioa	Erref
Eskuratu ditugun datuek (baita alor jakinetako adituek emandako iritziek ere) **adierazten** dute**nez**,	→	zientzia-alor jakin batean onartuko diren terminoak ebaluatzeko hierarkia bat ezarri behar da.	KAUSA	TERM18
Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unita[...]	←	**Izan ere**, iritzi ezberdinetako zientzialari serbiarrek adostasuna **lortu** dute eta aurreko hamarkadetan ingelesari eman diote zientzia-k [...]	KAUSA	TERM18
Alde batetik, gero eta indartsuagoa da nazioarteko harmonizazioa lortu beharra,	←	ekonomian, politikan eta kultura eta gizarte gaietan etenik gabe sortzen ari diren loturak eta elkarren arteko trukaketak **eraginda**;	KAUSA	TERM19
Terminologiak berak ere, uztartu egin behar ditu joera orokor horiek, eransten zaizkien beste batzuekin batera, hala nola: teknologien[...]	←	gizartearekin lotuta dagoen jarduera de**nez**,	KAUSA	TERM19

2.3. *Bilaketa aurreratuak*

Informazio morfosintaktikoa automatikoki aberastu ondoren, bilaketa aurreratuak egin daitezke hitz-forma, lema eta kategoria morfologikoetan oinarrituz. Bilaketa mota hori corpus osoan edota azpicorpus bakoitzean egin daiteke (http://ixa2.si.ehu.es/diskurtsoa/bilatzailea_mix/rst_treebank.php); horretaz gain, bilaketa unitate zentralera muga daiteke. Aipatu dugun bezala, bilaketa mota honen bitartez, bistara daitezke erlazio-seinale jakin baten adibide guztiak edo zehaztutako bi lemen agerpen guztiak (elkarren ondoan egon behar ez dutenak). Azkeneko kasu horren adibidea 6. taulan dugu. Adibidean ikus daiteke «talde» eta «helburu» lema duen hitz-formak zein esaldi tan dauden. 6. taulako bigarren zutabean testuaren erreferentzia agertzen da, hirugarren zutabean testuko zenbatgarren esaldian dagoen, laugarrenean, bilatutako elementuak agertzen diren hitzak, bosgarrenean unitate zentralean dagoen, seigarrenean bilatutako terminoak dauden esaldia bera. Horrelako bilaketek, adibidez, UZeko patroiak bilatzeko aukera ematen du.

6. TAULA

Euskal RST TreeBankeko datu-basean BILAKETAK atala

	Dok.	Sent Id	Hitza(k)	UZ	Esaldia
1	TERM50	sent2	taldeek / helburua	BAI	[...] Hitzaldi honek azken hiru urteotan lau unibertsitate hauen *talde*ek egindako ikerkuntzaren ondorioetako batzuk azaltzeko *helburua* izango luke.
2	ZTF13	sent1	taldearen / helburu	BAI	[…] Gure *ikerkuntza talde*aren *helburu* nagusia, […]
3	ZTF13	sent17	taldearen / helburu	EZ	Alor honetan, gure *ikerkuntza talde*aren *helburu* nagusiak bi dira.
4	ZTF17	sent1	talde / helburu	BAI	[…] *Talde* honek burututako lana materialen arloan kokatuta dago eta hiru konposatu-multzoen lorpena, karakterizazioa eta propietateen azterketa ditu *helburu*.
1	ZTF15	sent7	helburu / talde	EZ	[…] bestelako galdera zailagoei ere erantzutea dute *helburu*, hala nola, espezieen biogeografia, *talde*aren filogenia, eta abar.

3. **Ondorioak eta etorkizuneko lana**

Lan honetan RST eta Euskal RST *TreeBank*a aurkeztu ditugu. Horrez gain, aberastutako corpusa eskuragarri jarri dugu erabili nahi duen ororentzat. Egindako lana euskararen Hizkuntzaren Prozesamenduan egin behar diren zenbait atazatan erabilgarria da. Besteak beste, segmentazio automatikoa lortzeko urre-patroia lortu dugu.[9] Unitate zentralari dagokionez, testuen gai orokorra edo makroegituran dauden erregulartasunak bilatzeko eta detektatzeko bilaketa-tresna berritzailea inplementatu dugu. Erlazio erretorikoei dagokienez, erlazio-seinaleak etiketatzean erlazio erretorikoen patroiak diseinatu ahal izateko kontsulta sistema berritzailea sortu dugu. Horrela, beste hizkuntzetarako baliagarria diren kontsulta-zerbitzuak oraingoz RSTko bibliografian eskaintzen direnak baino osatuagoak eta hobeak egitea lortu dugu.

Bestalde, lan honen mugetako bat corpusaren tamaina da, beste corpus batzuekin konparatzen badugu, Euskal Treebanka txikiagoa da eta domeinu ezberdineko testuak badira ere, corpuseko testu guztiak testu-laburpenak dira; beraz, genero berekoak dira.

Etorkizunean hurrengo lanak egiteko asmoa dugu: i) erlazio erretorikoen patroiak definitu; ii) patroi horiek eta ereduzko corpusetik ateratako erlazioen probabilitatea integratu DiZer automatikoki diskurtsoa analizatzeko plataforma eleaniztunean (Pardo et al 2004);[10] iii) corpusa domeinu ezberdinetako testu-laburpen gehiagorekin osatzea eta beste genero batzuk etiketatzea; iv) erlazio-egiturako informazioa idazten laguntzeko tresnetan integratzea.

Aipamenak

Alberdi, Xabier & Julio Garcia. 2012. «Diccionario de marcadores discursivos del euskera (I)». *V Congreso Internacional de Lexicografia hispánica*. Madril.

Antonio, J.D. 2004. *Estrutura retórica e articulação de orações em narrativas orais e em narrativas escritas do português*. doktore-tesia, Faculdade de Ciências e Letras, U.E. Paulista, Brasil.

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge Univ Pr.

Bosma, Wauter E. 2008. *Discourse oriented summarization*. doktore-tesia. University of Twente.

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski. 2001. «Building a discoursetagged corpus in the framework of rhetorical structure theory». 2nd SIGDIAL, Aalborg, Denmark. 30-39.

---

[9] Diskurtso segmentatzailearen demoa hemen proba daiteke: [http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl]

[10] DiZer *analizatzaile diskurtsibo* eleaniztuna hemen kontsulta daiteke: [http://www.nilc.icmc.usp.br/dizer2/]

Carlson, Lynn, Mary E. Okurowski & Daniel Marcu. 2002. *RST Discourse Treebank*, LDC2002T07 [Corpus]. Philadelphia: PA: Linguistic Data Consortium.

Corston-Oliver, Simon. 1998. «Identifying the linguistic correlates of rhetorical relations». *ACL Workshop on Discourse Relations and Discourse Markers*. Canada. 8-14.

da Cunha, Iria, Juan-Manuel Torres-Moreno & Gerardo Sierra. 2011. «On the Development of the RST Spanish Treebank». *5th Linguistic annotation workshop* (LAW-5). AEB. 1-10.

da Cunha, Iria. Mikel Iruskieta. 2010. «Comparing rhetorical structures in different languages: The influence of translation strategies». *Discourse Studies* 12, 563-598.

Esnal, Pello. 2008. *Testu-antolatzaileen erabilera estrategikoa*. Bilbo: Euskaltzaindia.

Euskaltzaindia. 1990. *Euskal gramatika. Lehen urratsak III (Lokailuak)*. Bilbo: Euskaltzaindia.

Euskaltzaindia. 1994. *Euskal gramatika: lehen urratsak (EGLU) VI (juntagailuak)*. Bilbo: Euskaltzaindia.

Euskaltzaindia. 1999. *Euskal gramatika. Lehen urratsak-V (mendeko perpausak-I)*. Bilbo: Euskaltzaindia.

Euskaltzaindia. 2005. *Euskal gramatika. Lehen urratsak-VI (Mendeko perpausak-II)*. Bilbo: Euskaltzaindia.

Ghorbel, Hatem, Azfal Ballim & Giovanni Coray. 2001. «Rosetta: Rhetorical and semantic environment for text alignment». *Corpus Linguistics*. Lancaster University (UK). 224-233.

Goenaga, Iakes, Olatz Arregi, Klara Ceberio, Arantza Díaz de Ilarraza & Amane Jimeno. 2012. «Automatic Coreference Annotation in Basque». *11th International Workshop on Treebanks and Linguistic Theories*. Portugal. 115-126.

Hanneforth, T., S. Heintze & Manfred Stede. 2003. «Rhetorical parsing with underspecification and forests». *HLT-NAACL* 2003. Volume 2. AEB. 31-33.

Iruskieta, Mikel. 2014. *Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean*. doktore-tesia, UPV/EHU.

Iruskieta, Mikel, María J. Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi & Oier Lopez de la Calle. 2013a. «The RST Basque TreeBank: an online search interface to check rhetorical relations». *4th workshop RST and Discourse Studies*. Brasil. 40-49.

Iruskieta, Mikel, Iria Da Cunha & Maite Taboada. 2014. «A qualitative comparison method for rhetorical structures: Identifying diferent discourse structures in multilingual corpora». *Language Resources and Evaluation,* 1-47.

Iruskieta, Mikel & Iria da Cunha. 2010a. «El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera y español». *Calidoscopio* 8, 181-202.

Iruskieta, Mikel & Iria da Cunha. 2010b. «Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera». *XXVIII Congreso Internacional AESLA*. Vigo. 146-159.

Iruskieta Mikel, Arantza Díaz de Ilarraza & Mikel Lersundi. 2011a. «Bases para la implementación de un segmentador discursivo para el euskera». Anais do *III Workshop A RST e os Estudos do Texto* (18-29). Brasil. 18-29.

Iruskieta, Mikel, Arantza Diaz de Ilarraza & Mikel Lersundi. 2011b. «Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera». *Procesamiento del Lenguaje Natural* 47, 137-144.

Iruskieta Mikel, Arantza Diaz de Ilarraza & Mikel Lersundi. 2013b. «Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque». *Corpus Linguistics and Linguistic Theory*.

Iruskieta Mikel, Arantza Díaz de Ilarraza & Mikel Lersundi. 2014b. «The annotation of the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations». COLING 2014, the *25th International Conference on Computational Linguistics: Technical Papers.* Dublin. 466-475.

Larringan, Luis. 1995. *Testu-antolatzaileak bi testu motatan: testu informatiboa eta argudiapenezkoa*. doktore-tesia, UPV/EHU.

Lascarides, Alex & Nicholas Asher. 2007. «Segmented discourse representation theory: Dynamic semantics with discourse structure». *Computing meaning,* 87-124.

Mahmud, R. & A. Ramsay. 2005. «Finding Discourse Relations in Student Essays». *Computational Linguistics and Intelligent Text Processing*, 116-119.

Mann, Willian C. & Maite Taboada. 2010. RST web-site. [http://www.sfu.ca/rst/] (2012/04/24).

Mann, Willian C. & Sandra A. Thompson. 1987. «Rhetorical Structure Theory: A Theory of Text Organization». *Text* 8(3), 243-81.

Mann, Willian C. & Sandra A. Thompson. 1988. «Rhetorical Structure Theory: Toward a functional theory of text organization». *Text-Interdisciplinary Journal for the Study of Discourse* 8(3), 243-281.

Marcu, Daniel. 2000. *The theory and practice of discourse parsing and summarization*. Cambridge: The MIT press.

Mitkov, Ruslan. 2002. *Anaphora resolution*. Longman. London.

Oates, Sarah Louise. 1999. «State of the art report on discourse markers and relations». *Technical report*, University of Brighton, Information Technology Research Institute.

O'Donnell, Michael. 2000. «RSTTool 2.4: a markup tool for Rhetorical Structure Theory». *First international conference on Natural Language Generation* 14, 253-256.

Pardo, Thiago A.S. 2005. *Métodos para análise discursiva automática.* Instituto de Ciências Matemáticas e de Computação. doktore-tesia, Universidade de Sao Paulo.

Pardo, Thiago A.S., Maria G.V. Nunes & Lucia H.M. Rino. 2004. «DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese». 17th Brazilian Symposium on Artificial Intelligence-SBIA *Lecture Notes in Artificial Intelligence* 3171, 224-234.

Pardo, Thiago A.S. & Eloize R. M. Seno. 2005. «Rhetalho: um corpus de referência anotado retoricamente». *V Encontro de Corpora.* Brasil. 24-25.

Prasad, Rashmi, N. Dinesh, A. Lee, Eleni Miltsakaki, L. Robaldo, Aravind Joshi & Bonnie L. Webber. 2008. «The penn discourse treebank 2.0». *6th LREC*. Morocco. 2961-2968.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. A. Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio & Yannick Versley. 2010. «Semeval: Coreference resolution in multiple languages». *5th International Workshop on Semantic Evaluation*. 1-8.

Sainz, Matilde. (ed.). 2001. *Azalpenezko testu entziklopedikoaren azterketa eta didaktika*. Donostia: Erein.

Stede, Manfred. 2004. «The Potsdam Commentary Corpus». *Workshop on Discourse Annotation ACL*. AEB. 96-102.

Sumita, Kazuo, Kenji Ono, T. Chino & T. Ukita and S. Amano. 1992. «A discourse structure analyzer for Japanese text». *International Conference on Fifth Generation Computer Systems* vol.2. 1133-40.

Taboada, Maite & Jan Renkema. 2011. *Discourse Relations Reference Corpus*. [http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html] (2012/04/24).

Taboada, Maite & Debopam Das. 2013. «Annotation upon annotation: Adding signalling information to a corpus of discourse relations». *Dialogue and Discourse* 4, 249-281.

Taboada, Maite & Willian C. Mann. 2006. «Applications of Rhetorical Structure Theory». *Discourse Studies* 8, 567-588.

van der Vliet, Nynke, 2010a. Inter annotator agreement in discourse analysis. [http://www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/presentations/NvdV-Cohens-Kappa-2010.pdf] (2012/04/24).

van Dijk, Teun A. 1980. «The semantics and pragmatics of functional coherence in discourse. Speech act theory: Ten years later». *Versus* 26(27), 49-65.