Aproximació al Resum Automàtic per Marcadors Discursius

Irene Castellón Masalles i Laura Alonso Alemany

X-Tract WP 01/07

Índex

Capítol 1

Introducció

1.1 Motivació

Les noves tecnologies han possibilitat que un ampli sector de la població tingui accés a grans quantitats d'informació. Les conseqüències d'aquest fenomen són fonamentalment positives, ja que suposa el que s'ha anomenat com a 'democratització de l'accés a la informació' (Yushkiavitshus 1998). Més concretament, es facilita l'accés a dades que havien estat totalment invisibles o inabastables per limitacions purament tècniques, alhora que fomenta la divulgació d'informació de tot tipus i la creació de xarxes d'intercanvi de coneixements, en un sentit ampli del concepte. D'aquesta manera es redueix sensiblement un dels factors de desigualtats entre les persones, la possessió o no possessió de coneixements.

No obstant, l'accessibilitat a la informació té també un vessant negatiu. Tot sovint, les persones no són subjectes actius d'un accés selectiu a la informació, sinó que esdevenen objectes passius d'una exposició a grans quantitats de dades no desitjades, que requereixen un important esforç de processament suplementari, esforç que no pot ser utilitzat per a l'obtenció del coneixement realment desitjat. És així com la sobreinformació pot resultar en desinformació, ja que una barrera de dades indiscriminades o dirigides per interessos externs impedeix que cadascú accedeixi a la informació seguint els seus propis interessos o necessitats.

La gestió de la informació és un dels principals mitjans per superar la sobreinformació que es dóna en aquesta 'societat de la informació' en què vivim. Es pot tractar la informació, i per tant la gestió de la informació, des de molts punts de vista. Donat que la informació és un objecte complex que participa de molts aspectes de l'activitat humana, són moltes també les ciències que se n'ocupen: la semiòtica, la lingüística, la matemàtica, l'antropologia, la intel·ligència artificial, les ciències documentals, la psicologia, la filosofia i altres ciències que tracten de manera tangencial els sistemes de comunicació. Cadascuna d'aquestes ciències aporta perspectives, propostes i solucions per als problemes comuns en la gestió de la informació.

Però més concretament, quan la informació que es tracta és de tipus lingüístic, creiem que és apropiat abordar la gestió de la informació des d'una perspectiva també lingüística, ja que sembla l'aproximació més escaient per donar compte de la natura de l'objecte que es tracta, i, per tant, tractar-lo de manera més adequada. La presència de contingut lingüístic és especialment rellevant en les àrees que s'ocupen del processament dels textos, en què podem distingir objectius diversos, com extracció d'informació (EI), recuperació d'informació (RI), traducció automàtica (TA) o resum automàtic (RA). No obstant, algunes de les línies de processament del llenguatge natural (PLN) que treballen amb textos no paren gaire atenció a la informació lingüística que contenen, sinó que exploten tan sols el vessant matemàtic de la informació, o tan sols les característiques documentals dels textos. Això és així perquè el tractament de la informació de tipus lingüístic suposa un cost d'investigació i desenvolupament sovint difícil d'assumir.

La lingüística computacional (LC) intenta sistematitzar i formalitzar les característiques del llenguatge humà per al seu processament en sistemes computacionals. Creiem que aquesta perspectiva pot constituir la clau d'un salt qualitatiu en la millora dels mètodes i tècniques de PLN perquè incorpora un tipus de coneixement, el lingüístic, específicament adequat al tipus d'objecte que tracta. Com hem dit, el major inconvenient d'aquesta aproximació és que els sistemes basats en coneixement lingüístic tenen un important cost de desenvolupament i funcionament. No obstant, l'aportació de la dimensió lingüística és prou considerable com perquè s'hagin anat desenvolupant múltiples eines i recursos que, actualment, fan possible l'accés a informació lingüística dels textos de tipus i profunditat variable per les diverses llengües i aplicacions de PLN.

Com hem dit, l'aproximació lingüística resulta especialment valuosa en el tractament dels textos. Sembla evident que, per un abordatge satisfactori dels objectes textuals, cal tenir en compte les característiques que constitueixen un text com una unitat de sentit, i no com una col·lecció desestructurada d'unitats lingüístiques menors. Per tant, caldria analitzar i explicar els vincles de cohesió i coherència que es donen entre les diverses parts d'un text, i les relacions que estableix el text amb el seu entorn comunicatiu. Això suposa, però, que hem de poder donar compte de les unitats de què està format un text, bàsicament, les paraules i les oracions. Resoltes les qüestions que afecten aquestes unitats, es podria treballar en els fenòmens de tipus purament discursiu.

Malauradament, en l'estat actual de les tècniques de PLN, no podem dir que tots els nivells inferiors al text puguin ser tractats de manera satisfactòria per mitjans computacionals. Si bé és cert que un bon nombre de llengües disposen ja d'eines que permeten una anàlisi automàtica fiable de la informació morfològica, trobem importants mancances pel que fa a la sintaxi i la semàntica. En aquests nivells, sovint tan sols es poden donar solucions parcials, amb importants restriccions pel que fa al domini conceptual o al gènere documental que es tracta. Sembla, doncs, que les qüestions discursives resten inabordables, per tal com no disposem d'un coneixement prou ampli de les unitats que constitueixen el discurs. D'altra banda, sembla que el coneixement del món necessari per comprendre un text en les seves múltiples facetes també resta inabastable.

Però ja hem apuntat una possible solució per l'abordatge de la complexitat lingüística: una solució parcial. Si no disposem de prou informació per abordar la totalitat del text, podem optar per tractar aquells aspectes del text dels quals podem donar compte. Així doncs, la incorporació d'informació lingüística pot ser un mitjà d'enriquiment dels sistemes computacionals que es basen en mètodes estadístics. Aquest tipus d'informació sembla molt adequada per al tractament dels textos, i això, sembla evident, pot contribuir en gran mesura a millorar el seu processament computacional.

1.2 Objectiu de la investigació

Entre les aproximacions a la gestió de la informació de caire més lingüísitc, trobem especialment interessant el camp del resum automàtic: per una banda, la possibilitat de resumir sense esforç les grans quantitats d'informació amb què ens trobem bombardejats cada dia és un objectiu molt temptador. D'altra banda, el resum automàtic planteja qüestions que ens porten a considerar diferents perspectives d'estudi dels textos:

- els processos **cognitius** que en què es basa l'habilitat humana de resumir un text (Kintsch i van Dijk 1978)
- l'elaboració de resums en l'àrea de la **documentació** (Endres-Niggemeyer 1998)

– les claus **lingüístiques** que determinen com, a partir d'un text, es pot construir un resum que en conservi el contingut essencial (Sparck Jones 1993)

Sembla clar que una perspectiva interdisciplinar és especialment adequada per assolir un objectiu complex, el resum, a partir d'un objecte complex, el text: per abordar els textos, gestionar-los tenint en compte les seves diverses facetes, i obtenir-ne un resultat adequat al text original i als objectius proposats. No obstant, donat que una complexitat tal està més enllà de l'abast del nostre treball, ens hem proposat de centrar-nos en la vessant lingüística dels textos, per donar compte d'alguns dels aspectes que caracteritzen els textos com a objectes lingüístics i que ens permeten d'obtenir-ne de manera automàtica un resum que en conservi l'essència.

Efectivament, en tant que objectes lingüístics, els textos tenen característiques pròpies que els defineixen i els diferencien dels elements de la resta de nivells de la llengua. Podem englobar aquestes característiques en el concepte 'textura' de Halliday i Hasan (1976), que agrupa aquells fenòmens lingüístics que constitueixen un text com a un objecte cohesiu i cohesionat. Fenòmens com la referència, la substitució, l'el·lipsi, la conjunció i la cohesió lèxica exerceixen una funció de lligam entre les diverses parts d'un text, de manera que el constitueixen en un element unitari diferent d'una simple col·lecció d'expressions. Sembla clar que, per explicar adequadament un objecte textual, caldrà tenir molt en compte aquests factors, en la mesura en què les tècniques de PLN ho permetin.

Trobem que és especialment interessant de desenvolupar un sistema de resum automàtic dels textos que incorpori coneixement lingüístic específic per l'espanyol, per diverses raons: en primer lloc, no coneixem, fins al moment, cap sistema d'aquestes característiques en l'àmbit hispànic. Com ja hem dit abans, una eina d'aquest tipus ofereix unes prestacions que són del tot desitjables, alhora que constitueix un mitjà per aprofundir en fenòmens lingüístics poc estudiats i assajar maneres de tractar-los computacionalment. Fins ara, els sistemes de resum automàtic per l'espanyol han aplicat mètodes que no contemplen la dimensió lingüística dels textos, sinó que els consideren com a seqüències d'informació tractables de manera matemàtica, amb mètodes estadístics, o bé mètodes que exploten tan sols les característiques documentals dels textos. Creiem que la perspectiva lingüística pot ser molt enriquidora en aquest estat de coses.

En segon lloc, el PLN per la llengua espanyola ha assolit un grau de desenvolupament que fa possible un abordatge superficial del nivell lingüístic textual. Disposem ja d'eines que ens permeten realitzar una anàlisi automàtica dels textos en un nivell de complexitat bàsic (RILE 2000, González et al. 1995): analitzadors morfològics fiables (MACO (Arévalo et al. 2001), EMERGE (Santana et al. 2000), Zavrel i Daelemans 1999) i analitzadors sintàctics parcials de components sintagmàtics de baix nivell (Atserias et al. 1998b, González et al. 1995). També comencem a tenir informació sobre les relacions de tipus semàntic entre els mots, gràcies a recursos com la xarxa lèxico-semàntica EuroWordNet (Vossen et al. 1997) o lèxics computacionals (Civit et al. 2000). Un tipus de recurs cada cop més important en lingüística computacional són els corpus en format electrònic, anotats amb diversos tipus d'informació: morfosintàctica (Carmona et al. 1998), de subcategorització, (Monedero et al. 1995), de conceptes (Gonzalo et al. 1998), etc. No obstant, manca una resolució eficaç del nivell oració, ja que no s'ha aconseguit de tractar computacionalment de manera satisfactòria qüestions de sintaxi i semàntica oracional, com per exemple les dependències entre components oracionals o la representació del contingut proposicional.

Si considerem que la representació d'un text es construeix de manera incremental, la no resolució del nivell oracional suposa un greu inconvenient per a la resolució del nivell que podríem considerar com a immediatament superior, el text. Però, com hem apuntat abans, una solució parcial sembla no només possible, sinó també adequada: sovint, una gran complexitat dels models lingüístics de PLN ha suposat un impediment per a la seva implementació efectiva en aplicacions computacionals. Una aproximació més simple donarà informació més superficial sobre els textos, però, en canvi, tendirà a requerir un cost menor de desenvolupament i també un cost menor de processament. Una perspectiva d'aquest tipus pot resultar també molt enriquidora: potser abordant el nivell oracional alhora des de les paraules i grups sintagmàtics i des dels grups de significat textual, podrem contribuir a delimitar en certa mesura aquesta unitat intermitja, per ara insatisfactòriament explicada, que és la oració.

És el nostre objectiu dinal d'investigació, doncs, desenvolupar un sistema de resum automàtic per a la llengua espanyola, que tracti els textos tenint especialment en compte la seva natura lingüística. Creiem que el paràmetre de la llengua és determinant a l'hora de desenvolupar un sistema així, ja que marca de manera diferenciada i idiosincràtica les característiques dels seus objectes lingüístics, els textos . És per això que volem construir un sistema específic per les especificitats de l'espanyol. No obstant, creiem que és possible de sistematitzar i explotar per al resum automàtic característiques generals dels textos d'una llengua, sense restriccions de gèneres documentals o de dominis conceptuals. Si bé és cert que amb aquestes restriccions es pot aprofundir en la comprensió dels textos, considerem més interessant i més apropiat donar compte de les pautes generals que regeixen el nivell lingüístic textual.

Aquest objectiu generalista, juntament amb les nostres limitacions pel que fa a recursos i eines de PLN, ens han decantat per una aproximació superficial als textos. Com exposarem tot seguit, després de considerar amb atenció diverses línies de treball sobre resum automàtic, hem optat per intentar traslladar a l'espanyol una proposta que dóna resultats prou bons per a l'anglès i el japonès: l'anàlisi de l'estructura retòrica dels textos. Aquesta proposta es basa en el supòsit que és possible i rellevant de determinar, mitjançant marques formals de la representació superficial dels textos, les diverses parts de què es componen els textos i com es relacionen aquestes parts entre elles per formar un tot cohesiu i coherent. És per això que hem concebut un sistema de resum automàtic que, partint d'aquesta proposta, s'ajusta a les nostres possibilitats i intenta explotar la informació lingüística que tenim al nostre abast, per tal de donar una representació el més acurada i útil possible dels textos.

Per desenvolupar aquest sistema, cal delimitar i sistematitzar la informació lingüística en què es basa: les marques d'estructura dels textos, que hem anomenat *marcadors del discurs*. Hem intentat, doncs, determinar quines són les claus textuals que ens ajudaran a construir un resum adequat d'un document, i hem provat de formalitzar-les i sistematitzar-les per al seu ús en un sistema computacional. També hem desenvolupat una aplicació del coneixement lingüístic associat a aquestes marques en el tractament automatitzat dels textos: una primera versió d'una eina que, basant-se en aquestes marques, identifica en un text les unitats que el componen, i que hem anomenat *segments*.

En dur a terme aquestes tasques, hem observat algunes interaccions interessants entre diversos nivells lingüístics, així com possiblitats de col·laboració entre diverses aproximacions al resum automàtic dels textos. Esperem que, més endavant, puguem aprofundir en algunes intuïcions que apunten en el treball que presentem. Creiem que, per a un bon desenvolupament del sistema de resum que hem concebut, hauran de ser abordades i resoltes aquestes i altres qüestions d'interès per al resum automàtic, la lingüística computacional i la lingüística en general.

En **aquest treball** presentem, doncs, el treball realitzat en el darrer any, que s'estructura en tres parts: en primer lloc, un estat de la qüestió de les diverses aproximacions al resum automàtic i sobre els models computacionals del discurs que semblen més adequats per abordar-lo. En segon lloc presentem l'estudi que hem realitzat sobre l'ús dels marcadors del discurs en PLN. En tercer lloc, exposarem l'anàlisi i desenvolupament d'una eina informàtica basada en aquest estudi, amb l'objectiu de segmentar automàticament els textos espanyols per usos discursius: el *segmentador* retòric. Finalment presentarem algunes conclusions d'aquest treball i algunes línies d'aprofundiment i desenvolupament futures.

Capítol 2

Estat de la Qüestió

Ja des dels principis de la lingüística computacional i del processament del llenguatge natural, els camps de la lingüística que tracten els fenòmens especialment complexos, com la semàntica, el text o el discurs, han despertat un gran interès en els investigadors. No obstant, els primers treballs en aquestes àrees no passaven de sistemes de joguina que no tenien més projecció que la purament acadèmica, per tres raons bàsiques:

- les limitacions tècniques dels sistemes informàtics no permetien el tractament de quelcom tan complex com el llenguatge natural
- la intel·ligència artificial, encara una ciència molt jove, no s'havia desenvolupat fins al punt de proposar mètodes i teories que permetessin de compatibilitzar el llenguatge humà i els sistemes informàtics
- la lingüística, també una ciència relativament jove, encara no havia contemplat el llenguatge de manera prou formal per al seu tractament computacional

Amb el temps, s'han anat cobrint moltes de les mancances en aquestes tres àrees. És per això que no podem considerar desfasades moltes de les tècniques desenvolupades en les etapes més primerenques del resum automàtic. Actualment podem contemplar-los des d'un nou punt de vista, ja que el grau de desenvolupament actual de la ciència i la tecnologia permet concebre aplicacions d'aquestes tècniques per a usos reals.

D'altra banda, la tendència generalitzada en els sistemes de resum automàtic és la integració de múltiples paràmetres que col·laboren per obtenir una descripció més acurada i útil dels textos. Com ja hem dit, els textos són objectes complexos i, per tant, sembla que la manera més adequada de tractar-los és la que millor contempli aquesta complexitat. És per això que sembla especialment adequat integrar tots aquells paràmetres que puguin aportar informació rellevant sobre algun aspecte dels textos. Ara bé, resulta també molt convenient donar compte de la manera com s'integra aquest conjunt d'informacios heterogènies que podem obtenir dels textos amb els mitjans de PLN actuals: hem de tenir en compte que, amb tota probabilitat, els diversos tipus d'informació textual interactuen i contribueixen de manera col·laborativa a configurar els textos. Cal, doncs, donar compte d'aquestes interaccions per tal d'explotar de manera adequada la informació disponible.

En situar-nos en el PLN de l'espanyol, hem de tenir en compte un paràmetre addicional, sovint obviat en el camp del resum automàtic: la llengua. En PLN, la major part d'esforços s'han dedicat al desenvolupament de la llengua preeminent en l'àmbit tecnològic, l'anglès. És cert que totes les llengües comparteixen característiques comunes, per tant, podem considerar que una part de la investigació en PLN és independent de llengua, però és igualment cert que existeix una gran variabilitat entre les diverses llengües, i que cal donar compte d'aquestes diferències de forma individual (Santos 1999). El paràmetre de la llengua s'ha de tenir en compte en la revisió bibliogràfica que presentem tot seguit. Donat que els treballs més destacats en l'àrea de resum automàtic són aplicacions per a l'anglès, hem de considerar la seva aportació al PLN i a la caracterització dels fenòmens lingüístics generals, però hem de tenir en compte que qualsevol proposta haurà de ser replantejada i adaptada per a l'espanyol.

El següent apartat intenta descriure les grans línies de treball en resum automàtic. Seguidament exposem amb més detall la línia que ens sembla més interessant d'aprofundir, l'anàlisi del nivell discursiu dels textos i, més concretament, de les relacions que podem distingir entre les diverses parts d'un text, i, finalment, esbossem una proposta per a una aproximació al resum automàtic (RA).

2.1 Estat de la qüestió en Resum Automàtic

Tot seguit presentem una revisió de l'estat de la qüestió en RA, distingint tres grans línies de treball:

- una primera aproximació que explota els aspectes **quantitatius** dels textos, considerant-los des del punt de vista dels mètodes estadístics propis de la recuperació d'informació (RI)
- una segona perspectiva que tracta d'explotar la informació **estructural** dels textos, a través de característiques com el format dels documents, la distribució espacial dels fragments més rellevants d'un text o l'ús de paraules clau de tema o d'estructura del document
- una darrera via que tracta de realitzar una **comprensió profunda** dels textos, aplicant tècniques de l'extracció d'informació (EI), és a dir, que intenta trobar en un text, habitualment d'un domini restringit, la informació estipulada a priori en un esquema que modelitza un cert coneixement del món, tot i que també poden aplicar-se les tècniques d'EI a textos irrestrictes per extreure tipus específics d'informació.

Hem abordat cadascuna d'aquestes línies amb una certa perspectiva històrica, per tal de reflectir els progressius avenços que s'han assolit en RA. No obstant, també hem intentat donar compte de les principals aportacions i limitacions de cada aproximació, independentment de la seva situació en el temps. D'aquesta manera podrem arribar a determinar quina de les aproximacions respon de manera més adequada a les nostres necessitats i objectius de PLN.

Classificar els diversos sistemes de resum automàtic no és una tasca fàcil, ja que sovint incorporen múltiples tipus d'informació textual i combinen diverses aproximacions. D'aquesta manera hom pretén donar compte de la natura complexa de l'objecte que es tracta, el text. En aquest estat de la qüestió, hem optat per distingir entre les tres línies de treball anteriors, però també hem tingut en compte, de manera transversal, altres paràmetres que caracteritzen les diverses aproximacions al resum automàtic, com per exemple:

- **extracte o resum**: segons si un sistema de resum tracta determinar quines són les parts més rellevants d'un text per tal d'extreure-les i confegir el resum encadenant aquestes parts literals del text, o bé si el que es pretén és construir de bell nou un text més curt que l'original que en reflecteixi els continguts més importants. Creiem que aquest paràmetre està molt determinat per les eines de PLN de què disposen els desenvolupadors d'un sistema, més concretament, si disposen d'eines de generació de llenguatge natural. Donat que la generació de llenguatge natural suposa una complexitat difícil d'assumir, ara com ara, la major part de sistemes practiquen un resum de tipus extracte. Tan sols el sistema de Radev i McKeown (1998) sembla disposar de la tecnologia necessària per elaborar un resum autònom, però fins i tot en el marc d'aquest grup d'investigació s'estudien alternatives menys costoses a la generació de llenguatge, com és l'aproximació híbrida que presenta Jing (1998).

- nivell de **compressió:** longitud del resum respecte del text original. Es distingeix entre les aproximacions que tracten d'obtenir un resum amb una longitud predeterminada (el 10% del text original, una única oració, la capacitat d'una pantalla d'ordinador, etc.) i els que donen un resum en què la informació més rellevant del text original es troba recollida en el mínim de volum possible, sense que aquest volum estigui prèviament establert. En general, la major part de sistemes assagen totes dues opcions.

- nivell de **retenció**: percentatge d'informació del text original que es recull en el resum. Per tal de determinar el nivell de retenció es fan servir mètodes estadístics i enquestes a humans. En el primer cas, es contrasta el resum obtingut amb un model ideal o bé amb paraules clau del text, i es calcula la cobertura i la precisió del resum automàtic respecte de l'ideal. En el segon cas, es contrasten les respostes a un mateix qüestionari de dos grups d'humans: uns que han llegit el text original i uns altres que n'han llegit el resum. La informativitat del resum es calcula a través del nombre de respostes correctes dels humans que no han llegit el text original.

- **dependent o independent** de gènere i/o domini conceptual: com a norma general, la dependència de gènere i/o domini és directament proporcional a la quantitat de coneixement incorporat al sistema: les aproximacions que aborden el resum automàtic mitjançant coneixement, ja sigui lingüístic o documental, solen delimitar la variació possible dels textos per tal d'aprofundir en la seva caracterització, és a dir, posen més èmfasi en la profunditat del coneixement que utilitzen que no pas en l'extensió de textos que tracten. Això implica que, sovint, els sistemes que incorporen coneixement treballen de manera efectiva amb fortes restriccions de gènere o domini. En canvi, els sistemes que exploten la informació quantitativa dels textos no solen presentar aquest tipus de restriccions, ja que prioritzen l'extensió dels textos que tracten per sobre de la profunditat amb què ho fan.

- **indicatiu o orientat a pregunta**: com hem dit, els sistemes de resum automàtic pretenen determinar els elements més rellevants d'un text per tal de construir el resum a partir d'aquests elements. La major part de sistemes intenten trobar claus per determinar, en un text, quines són les parts més rellevants. També trobem, però, alguns sistemes que determinen la rellevància dels elements d'un text segons si estan més o menys estretament relacionats amb una pregunta (Carbonell i Goldstein 1998) o un tema requerits per un usuari (Chali et al. 1999). A l'hora d'avaluar els sistemes de resum automàtic, en el concurs SUMMAC (Mani et al. 1998) s'ha proposat que s'haurien d'avaluar de manera diferenciada aquests dos tipus de tasca, ja que sembla que la darrera hagi d'esdevenir cada cop més i més important en PLN, a mesura que els avenços tecnològics puguin garantir-ne la fiabilitat.

- **multidocument o un únic document**: alguns sistemes intenten dur a terme resums de múltiples documents relacionats entre ells per algun criteri comú. Aquesta aproximació al resum requereix un tractament aprofundit dels textos, que necessàriament ha d'incorporar una important quantitat de coneixement. Hi ha diverses alternatives per tal d'assolir una representació satisfactòria sobre algun o diversos aspectes dels textos. Així, per exemple, Barzilay et al. (1999) incorporen tècniques d'EI, Ando et al. (2000) intenten explotar la informació temàtica, Carbonell i Goldstein (1998) proposen la mesura MMR[1] per detectar els segments de text que aporten informació diferent a la que el resum ja conté. Altres autors treballen sobre un vessant més matemàtic de la informació textual: Mani i Bloedorn (1997) utilitzen grafs, i Radev et al. (2000b) utilitzen representacions vectorials combinades amb tècniques de detecció i seguiment de tema.

- no podem oblidar que cada **llengua** presenta característiques pròpies que han de ser abordades de manera específica (Santos 1999). És per això que els sistemes de resum solen desenvolupar-se per una llengua concreta. Alguns sistemes, però, són capaços de tractar diverses llengües perquè no incorporen coneixement lingüístic de cap tipus, sinó que tan sols

[1] Maximal Marginal Relevance (rellevància marginal màxima)

tracten els aspectes quantitatius dels textos o bé la seva configuració en tant que documents. No obstant, els sistemes capaços de tractar diverses llengües que incorporen coneixement lingüístic solen desenvolupar mòduls específics que contemplen les particularitats de cadascuna (Hovy i Lin 1998).

– **format** d'entrada o sortida: el suport en què es dóna el resum final determina moltes de les característiques d'aquest resum: si és un resum parlat, per exemple, tendirà a incorporar frases curtes, si ha d'adaptar-se a una mida concreta, haurà de desenvolupar heurístiques per reduir el volum d'un text sense perdre qualitat en el contingut, etc. De la mateixa manera, serà molt diferent fer resums de textos escrits.

– en el plantejament del resum automàtic també es contempla la possibilitat d'elaborar resums **orientats a tipus d'usuaris** (Mani i Bloedorn 1998), amb unes necessitats d'informació concretes i uns requeriments molt definits pel que fa a longitud, registre lingüístic, llengua final del resum, etc. Aquest paràmetre és especialment determinant en els sistemes de resum desenvolupats per a la gestió de la informació en grans grups empresarials.

Hem intentat tenir en compte tots aquests paràmetres a l'hora d'estudiar i sistematitzar les diverses aproximacions al resum automàtic. No obstant, ens hem centrat en aquelles que ens han semblat més adequades als nostres objectius i limitacions: per una banda, ens interessen especialment les que tracten els aspectes més generals dels textos, perquè tenen un abast més ample i creiem que poden ser més explicatives en un nivell lingüístic. Així, hem parat especial atenció a les aproximacions sense restriccions de gènere o domini conceptual que tracten de fer resums informatius. No obstant, hem tingut sempre en compte una restricció que considerem determinant: la llengua. D'altra banda, trobem que és adequat de tractar els textos mitjançant la incorporació de coneixement, per la qual cosa pararem especial atenció a les aproximacions que fan contribucions significatives en aquest sentit. Hem de tenir en compte, doncs, que l'estat de la qüestió que presentem tot seguit està certament dirigit pels interessos que acabem d'exposar.

Malgrat aquesta abundància de paràmetres classificatoris, en aquest estudi n'hem trobat a faltar un d'essencial: una referència sòlida i consensuada pel que fa a l'**avaluació** de les diferents aproximacions al resum automàtic. Tal com apunta Radev (2000a), els mètodes d'avaluació de resum automàtic de què disposem fins ara són o bé massa estadístics o bé massa humans, de manera que mai no es dóna un resultat prou objectiu i representatiu alhora de la qualitat dels resums. A més, hem de tenir en compte que els resums són objectes polifacètics i poden ser molt diferents entre ells, de manera que l'avaluació pot variar significativament depenent de les diferents tasques o els tipus de resum. Ara bé, si restringim algun dels paràmetres que caracteritzen els resums, simplificarem aquesta complexitat i serà possible fer una avaluació més objectiva i precisa dels diversos sistemes de RA. Per exemple, Jing (1998) apunta que, si s'especifica una longitud determinada pels resums, s'obtenen mesures de cobertura i precisió dels resums molt més significatives.

En els treballs sobre resum automàtic que es publiquen habitualment, l'avaluació dels resultats no es fa segons uns paràmetres consensuats, sinó que cada investigador dóna compte de les aportacions que considera més rellevants del seu sistema. En termes generals, aquestes aportacions tracten de solucionar alguna de les principals mancances que presenten els sistemes de resum automàtic actuals. Tal com ho presenta Paice (1990), les mancances dels sistemes de resum poden agrupar-se en:

– manca de **representativitat** respecte del text original, és a dir, que no es reflecteixen les parts més importants del text. Paice proposa l'anàlisi retòrica del text per tractar de solucionar aquesta manca d'adequació.

– manca de **coherència**. Sovint, en extreure fragments, es trenquen els vincles referencials que constitueixen l'entramat del text. Per obtenir un resum coherent, caldrà donar compte de l'anàfora en el text original. Paice distingeix tres tipus d'anàfora a resoldre: pronominal, lèxica i de connexió retòrica.

Molts sistemes de resum destaquen també, com a part significativa de la seva avaluació, les solucions que aporten a les limitacions tècniques del resum automàtic, tal com les acabem de presentar (resum de múltiples documents, independència de domini, etc.).

El concurs SUMMAC (Mani et al. 1998) tracta d'establir un consens respecte de l'avaluació dels sistemes de resum automàtic, i per tant proposa un mètode d'avaluació que intenta tenir en compte els diversos aspectes dels resums. En aquest concurs s'avaluen 16 sistemes a través de tres tipus de resum diferents: un resum informatiu de longitud fixa (el 10% del text), un de longitud variable i un que ha de respondre a un tema determinat. Aquestes tasques s'han de realitzar sobre un corpus comú, procedent del concurs TREC (Harman i Voorhees 1996), que consta de 20 temes amb 50 documents diferents cadascun. La longitud dels documents va de 2.000 a 20.000 caràcters.

La bondat dels resultats de cada sistema s'avalua amb mesures extrínseques, més objectives, però també intrínseques, per tal de recollir la qualitat del resum en tant que text. Les mesures intrínseques avaluen qüestions com el fet que es recullin les principals idees d'un text (Paice 1990), la fluidesa del resum (Brandow et al. 1994) o la similaritat amb un resum ideal (Edmunson 1969, Kupiec et al. 1995). Algunes d'aquestes mesures poden obtenir-se mitjançant mesures estadístiques com la precisió o la cobertura, però per calcular-ne d'altres cal recórrer a jutges humans. Ara bé, els humans solen mostrar grans disparitats de criteri entre ells, de manera que costa establir una referència comuna clara amb què contrastar els resums automàtics. En canvi, les mesures extrínseques superen aquestes disparitats, i resulten, per tant, molt més objectives. Aquest tipus d'avaluació se sol basar en els resultats obtinguts per cada sistema en una tasca determinada: per exemple, l'efectivitat d'un resum per determinar si un text pot ser categoritzat en un determinat tema, o la informativitat d'un resum mesurada a partir de les respostes que donen jutges humans a un qüestionari.

Els resultats obtinguts en el SUMMAC pels tres tipus de resum de cada sistema en cada tasca es contrasten amb els resultats de les mateixes tasques utilitzant el text original complet i un resum que recull el primer 10% del text. Sorprenentment, aquests darrers obtenen resultats considerablement bons, sobretot si els comparem amb els resultats dels resums de longitud fixa. Sembla que aquest fet es pot generalitzar per tots els textos de tipus periodístic de menys de 1500 paraules (Wasson 1998, Brandow et al. 1994). En canvi, els resums de longitud variable solen obtenir resultats tan bons com els que s'obtenen utilitzant el text complet. No obstant, la informació obtinguda dels informants humans no queda satisfactòriament sistematitzada, de manera que resulta informativa però no determinant.

En general, trobem que els resultats d'aquest concurs són indicatius, però insuficients. Per una banda, no s'ha aconseguit d'establir un ideal de resum amb què es puguin contrastar els resultats obtinguts pels diversos sistemes. Per altra banda, les tasques d'avaluació proposades no arriben copsar tota la complexitat dels resums, de manera que l'avaluació resulta parcial. Alguns autors proposen estratègies d'avaluació alternatives: per exemple, Goldstein (1999) es basa en avaluació extrínseca, però proposa tasques diferents de les del SUMMAC. Radev (2000b) calcula la utilitat de cada oració en un conjunt d'oracions similars, mentre que Marcu (1999) i Jing et al. (1998) determinen les oracions més importants d'un text, que haurien d'estar incloses en un resum. No obstant, trobem que manca un punt de referència pel que fa a l'avaluació dels sistemes de resum automàtic, que pugui orientar aproximacions millorades al resum automàtic i permeti valorar en la seva justa mesura les aportacions dels sistemes actuals.

2.1.1 Aproximacions Quantitatives

En un estadi molt primerenc del desenvolupament de la intel·ligència artificial, trobem un primer sistema de resum que explota les possibilitats tecnològiques del moment: el sistema de Luhn (1958) determina quines són les paraules més freqüents d'un document[2], detecta aquelles frases en què ocorren més paraules freqüents i construeix un resum del text concatenant aquestes frases una darrera l'altra, seguint el seu ordre d'aparició en el document original. D'aquesta manera s'encetava una de les línies d'investigació en resum automàtic que més s'ha treballat.

Malgrat que el sistema de Luhn pot semblar molt simple, encara avui la **freqüència de paraules** és un dels paràmetres més utilitzats en els sistemes de resum automàtic. Val a dir, però, que les mesures de freqüència de mots s'han millorat des d'aleshores, sobretot a partir dels treballs en recuperació d'informació (RI). La mesura de freqüència més utilitzada actualment és la *tf*idf*[3] , que pondera la freqüència d'un terme en un document en relació a la freqüència inversa del mateix terme en una col·lecció de documents, de manera que es troben els mots freqüents en un document alhora que poc freqüents en la resta de documents, és a dir, els mots que singularitzen, per tant, caracteritzen, un document. D'aquesta manera s'elimina el que s'anomena 'soroll', és a dir, la informació no rellevant per a la tasca proposada, ja sigui recuperació d'informació responent a una cerca, resposta de preguntes o resum automàtic.

Una vessant més lingüística del mètode *tf*idf* són les **llistes de paraules buides**: en cada llengua es donen un conjunt de paraules funcionals, sense contingut lèxic; aquestes paraules no contribueixen en absolut al contingut d'un document, però són molt freqüents, cosa que pot esbiaixar el càlcul per freqüència de les paraules més rellevants d'un document, de manera que no quedi reflectit en absolut el seu contingut. Per evitar-ho, solen utilitzar-se les anomenades 'llistes de paraules buides': donat que la major part d'aquestes paraules pertanyen a categories tancades (preposicions, conjuncions, pronoms), es poden definir per extensió, és a dir, es poden tractar en llistes. Un mètode habitual per a determinar les paraules més rellevants d'un document, reduïnt al mínim el 'soroll', consisteix a ignorar en primer lloc les paraules funcionals contingudes en llistes, i aplicar sobre les paraules restants la mesura *tf*idf*. La major part de sistemes comercials utilitzen variants d'aquests tipus de mesures.

Lin (1996) intenta anar una mica més enllà: proposa de computar, ja no la freqüència de paraules, sinó la **freqüència dels conceptes** que es donen en un text. Per fer-ho, utilitza una xarxa lèxico-semàntica, WordNet (Miller et al. 1990), per generalitzar els conceptes d'un text a partir de les paraules particulars que hi apareixen. Lin intenta assolir un grau de generalització òptim en la xarxa lèxico-semàntica, és a dir, ni massa concret ni massa general, mitjançant una mesura anomenada 'interesting wavefront', que té en compte paràmetres com la distància entre els conceptes en la xarxa i entre els conceptes que apareixen en el text. Aquest tipus d'aproximació també ha estat aplicat per a l'espanyol (Gonzalo et al. 1998) en l'àmbit de la recuperació d'informació. En contra del que podríem esperar, en cap d'aquests dos camps els resultats obtinguts amb una mesura de freqüència de conceptes suposen una millora especialment significativa respecte de les mesures de freqüència de paraules. Probablement, una adaptació de WordNet específicament per tasques d'aquest tipus podria suposar un salt qualitatiu en l'ús de les freqüències de conceptes per al PLN. Hem de dir que Lin no fa cap referència a la desambiguació del sentit de les paraules ambigües (Gonzalo et al. 1999). Sense resoldre aquesta qüestió, no és possible d'aplicar el mètode de freqüències de conceptes, ja que resultarà impossible determinar a quin concepte s'està referint una paraula ambigua.

[2]per determinar a partir de quin nivell de freqüència una paraula és important en un document, Luhn va realitzar experiments amb informants humans.

[3]term frequency / inverse document frequency

També cal tenir en compte el treball de Hovy i Lin (1998), que fa una abstracció del mètode de freqüència de paraules aplicat a tècniques de classificació de documents. Per abordar el resum d'un document no assignat a cap domini, els autors el contrasten amb els conjunts de les paraules més freqüents, per tant, més característiques, d'uns dominis determinats. Si entre les paraules més rellevants del document es troben les paraules que caracteritzen un cert domini, el document és assignat a aquest domini. Tant les paraules més rellevants d'un document com les d'un domini són determinades mitjançant la mesura de freqüència de paraules *tf*idf* i les llistes de paraules buides. No obstant, a diferència de les tècniques de classificació de documents, aquest mètode implica un coneixement previ important dels dominis en què ha de treballar el sistema, ja que la construcció dels **conjunts de paraules característiques per cada domini** s'obté per tècniques d'aprenentatge automàtic entrenades a partir de conjunts de documents que uns especialistes humans han classificat prèviament per dominis. Aquest treball es porta a terme en el marc d'un sistema de resum automàtic que incorpora diversos tipus d'informació textual, SUMMARIST (Hovy i Lin 1998): un cop detectat el domini al qual pertanyen els textos, s'apliquen estratègies de resum adaptades a cada domini particular. Aquestes estratègies contemplen informació de freqüència de paraules però també integren la dimensió documental dels textos, com presentarem més endavant (2.1.2)

Els principals avenços en resum automàtic se centren sobretot en sistemes desenvolupats específicament per l'anglès, mentre que la resta de llengües estan força desateses, per la qual cosa no disposem de gaire evidència de com resulta millor d'abordar les particularitats de les diverses llengües. En l'àmbit lingüístic romànic, podem destacar **un sistema per al francès** amb un funcionament semblant al que presenten Hovy i Lin (1998), el de Lehmam (1995, 2001). Aquest sistema treballa també a partir d'un conjunt de paraules específic per una sèrie de gèneres i dominis particulars (s'han treballat dominis com economia, química, genètica, etc.). No obstant, intenta anar una mica més enllà del que proposa SUMMARIST, ja que no només treballa amb conjunts de paraules rellevants per un domini concret, sinó que construeix una base de dades que conté un conjunt de mots considerats com a rellevants en qualsevol document, els anomenats 'marcadors lingüístics d'extracció'. Malgrat la semblança en el nom, aquests 'marcadors lingüístics' exploten un tipus d'informació diferent que els 'marcadors del discurs'. Com exposarem en el següent apartat, els primers intenten caracteritzar un text pel seu contingut, mentre que els segons volen donar compte de la seva estructura.

Trobem que el principal desavantatge d'aquesta proposta rau en la metodologia emprada per al seu desenvolupament: donat que els conjunts de 'marcadors lingüístics' han estat elaborats a mà, el desenvolupament del sistema comporta un cost molt alt, i, alhora, és molt possible que el seu abast sigui més limitat que el dels sistemes que parteixen de mesures estadístiques. Seria interessant d'estudiar els resultats d'aplicar mètodes automàtics a l'adquisició i organització del coneixement lingüístic necessari per al funcionament d'aquest sistema.

A més, el sistema de Lehmam presenta una altra diferència considerable amb els sistemes més propis de l'anglès: es treu partit d'informació morfològica, és a dir, que es du a terme una anàlisi morfològica del text a resumir. Aquest tipus de procés no és gaire habitual en l'àmbit anglòfon perquè la flexió de les paraules angleses és molt escassa i, per tant, la informació morfològica associada als mots és poc informativa. En canvi, les llengües romàniques solen presentar una morfologia molt rica que pot aportar informació molt útil sobre els mots i els textos. Així, Lehmam sotmet els 'marcadors lingüístics' generals o dependents de domini a extensions morfològiques (nombre, persona, gènere) i lèxiques (derivació, sinonímia), per tal d'ampliar l'abast del sistema, ampliant el nombre de de marcadors lingüístics que es poden reconèixer. Així mateix, la informació morfològica contribueix a determinar quins fragments del text són més rellevants: per exemple, tan sols s'extreuen aquelles oracions en què el verb no està en cap forma del subjuntiu i si la persona és la primera o la tercera. Creiem que la proposta de Lehmam presenta de manera molt significativa com la idiosincràsia de les llengües romàniques, lluny de constituir un desavantatge a l'hora d'adaptar els sistemes proposats per a l'anglès, pot ser explotada com a font d'informació lingüística per aconseguir uns millors resultats en la construcció d'un resum.

Els mètodes de **cadenes lèxiques** són una altra evolució interessant dels mètodes que exploten la freqüència de paraules pel resum dels textos, ja que tracten els valors purament estadístics aportant-los estructura lingüística, ja que tracten d'explotar els patrons de distribució de les paraules en els textos (Barzilay i Elhadad 1997, Fuertes 2001). La principal diferència amb les mesures de freqüències tradicionals rau en el fet que no es computen els ítems lèxics aïllats, com una 'bossa de paraules' sense estructura, sinó que es considera que les paraules estan relacionades entre elles per similaritat lèxica. Aquesta similaritat pot determinar-se a partir d'una xarxa semàntica, d'un tesaure o senzillament tenint en compte la repetició (total o parcial) dels termes en un text. Detectant quines paraules estan relacionades entre elles, es poden detectar cadenes de termes que representen el fluxe d'informació en el text, de manera que cada cadena es correspondria amb una línia temàtica. A partir d'aquesta informació explícita sobre l'organització de la informació en el text, es pot determinar que les oracions més importants d'un text seran les més representatives de les línies temàtiques, és a dir:

- les que continguin termes de més cadenes lèxiques diferents
- les que continguin els termes més importants (normalment, els primers) d'una cadena lèxica
- les que continguin termes de les cadenes lèxiques més importants (normalment, les cadenes més presents en el text per quantitat de termes o per extensió)

En les aproximacions al resum automàtic per cadenes lèxiques, podem observar una certa gradació en la profunditat de processament del llenguatge. Així, trobem des dels autors que es basen en la simple repetició dels termes (Boguraev i Neff 2000b) fins aquells que van més enllà de la similaritat lèxica tal com la treballen Barzilay i Elhadad i intenten copsar fenòmens de **co-referència** (Boguraev i Kennedy 1997, Azzam et al. 1999). Salton et al. (1997) apliquen la teoria d'**hipervincles** a les cadenes lèxiques, i obtenen una estructura del document basada en el que anomenen 'hipervincles semàntics'. Aquests hipervincles reflecteixen una estructura del text en què es poden identificar blocs i relacions temàtics, de manera que es pot determinar que els vincles o cadenes lèxiques entre parts del text altament connectades seran els més representatius del contingut del text.

Sembla, doncs, que les aproximacions quantitatives als textos no suposen necessàriament un tractament superficial del material lingüístic, sinó que poden establir una estreta relació amb tècniques considerades més profundes, com l'Extracció d'Informació. En efecte, la proposta d'Azzam et al. (1999) està basada en la resolució d'anàfora que proporciona el sistema d'EI LaSIE (Humphreys et al. 1998). En aquestes darreres aproximacions al resum automàtic es realitza una comprensió del text que les fa molt properes a la tercera línia de treball que presentem, la de la comprensió profunda dels textos.

Probablement, la profunditat d'explicació dels textos que podem assolir mitjançant mètodes quantitatius està estretament relacionada amb el **tractament matemàtic de la informació**. Tal com ja apunta Shannon (1951), podem donar compte amb mètodes matemàtics de certa estructura de la informació. En el seu treball, Shannon determina que es dóna una mitjana d'un 75% de redundància en la llengua anglesa. Així doncs, podem arribar a una considerable profunditat explicativa dels textos si explotem la dimensió matemàtica de la informació que contenen. No obstant, això no ens pot fer oblidar que els textos no són objectes purament matemàtics, sinó que cal donar compte igualment de les seves altres facetes: la lingüística, la documental, la més semiòtica, etc.

2.1.2 Informació Estructural

Malgrat que el treball de Luhn va obrir tota una nova àrea d'investigació, l'estat de desenvolupament de la tecnologia de computació de l'època el rellegava a la condició de sistema de joguina, de manera que en el seu moment tan sols va despertar interès com a especulació sobre unes possibilitats futures que semblaven molt llunyanes. És per això que no és fins una dècada després del treball de Luhn que va sorgir una altra aportació significativa en aquesta àrea: el sistema d'Edmunson (1969).

Per realitzar un resum d'un text anglès bast en l'extracció de les oracions més rellevants, Edmunson combina altres mètodes amb les tècniques de freqüència de paraules: és el primer a sistematitzar l'aportació d'algunes **paraules clau** que indiquen la importància de certes parts del text (*en conclusió*, *com a síntesi*,...) o la manca d'importància d'altres parts del text (*per exemple*,...). També considera la **posició** de les paraules en el text per determinar-ne la rellevància: postula que aquelles paraules que ocorren en posicions preeminents, com ara els títols o subtítols, o els inicis de paràgraf, tenen més importància que la resta.

Així, Edmunson té en compte la informació que contenen els textos en tant que estructures d'informació. Podem distingir dos vessants en la dimensió estructural dels textos: la lingüística i la documental. Efectivament, una hipòtesi àmpliament acceptada (Halliday i Hasan 1976, Hobbs 1979, Dressler i Beaugrande 1981, Grosz i Sidner 1986, Polanyi 1988) és que els textos, en tant que objectes **lingüístics**, estan formats de parts. Les relacions entre aquestes parts constitueixen el text com a unitat lingüística diferenciada de la resta. Sembla que podem donar compte d'aquest entramat de parts i relacions com a una estructura. Donat que els textos s'articulen en diversos nivells de significat, podem distingir diverses estructures en un mateix text, segons el tipus d'informació lingüística que reflecteixen: l'estructura intencional, l'estructura ideacional, l'estructura textual, l'estructura argumentativa, l'estructura d'informació, etc. Podem recórrer a diversos mecanismes per explotar la informació de cadascun d'aquests nivells. Edmunson, per exemple, proposa d'explotar algunes marques lingüístiques que donen compte de certa estructura argumentativa dels textos, les paraules clau.

Alhora, els textos també tenen tenen una estructura determinada per la seva dimensió **documental**. Les ciències documentals han tractat de donar compte de l'aportació que poden fer qüestions relacionades amb el format dels documents per a la caracterització del seu contingut, tenint en compte paràmetres com el gènere o el domini. Així, es pot sistematitzar la manera com la informació que trobem en els títols i subtítols caracteritza el contingut d'un document, o es pot treure partit de la distribució d'un llibre en capítols per reconèixer els principals temes i delimitar els blocs temàtics en què s'articula el text. Així doncs, la informació d'estructura d'un document constitueix també una clau molt valuosa per descobrir l'organització de la informació en els textos. Edmunson tracta d'explotar aquest tipus d'informació considerant les parts d'un text en relació a la seva posició en el format del document.

Aquestes primeres observacions sobre el paper de les paraules clau i de la posició en la distribució de la informació en els textos s'han desenvolupat molt fins al moment actual; constitueixen una peça cabdal del funcionament de la major part de sistemes de resum automàtic comercials, i en, general, en molts sistemes que prioritzen la velocitat d'anàlisi per sobre de la profunditat (per exemple, els sistemes de resum automàtics incorporats en un motor de cerca a Internet) o els qui no tenen al seu abast eines de PLN que els permetin una anàlisi profunda dels textos (per exemple, per llengües amb pocs parlants). Ara bé: una de les principals mancances d'aquests mètodes és que solen ser molt dependents de gènere. Si podem restringir el tipus de text amb què treballem, és possible sistematitzar la informació estructural d'un document i extreure'n informació rellevant. Però si hem de tractar textos sense cap restricció, resulta molt difícil explotar aquest tipus d'informació, ja que no podem sistematitzar l'aportació de marques d'estructura del document per caracteritzar el contingut del text. Les característiques estructurals poden tenir un comportament molt diferent en tipus de text diferents. Així, per exemple, algunes paraules clau poden explicitar una determinada importància o organització de la informació en notícies periodístiques, i una altra de ben diferent en informes mèdics. El mateix s'aplica a les posicions relatives de les unitats lingüístiques en els documents. Sembla, doncs, que aquestes característiques són útils per tractar textos controlats, en dominis restringits, però no sembla que puguin donar compte de tot tipus de textos.

En el sistema de resum automàtic SUMMARIST (Hovy i Lin 1998) s'ha desenvolupat un mòdul que explota la informació que aporta el format dels documents, en especial, la **posició** relativa de les oracions, per determinar quines són les parts o les paraules més rellevants d'un text. En aquest sistema, diversos tipus d'informació contribueixen a determinar de manera col·laborativa les oracions més importants del document: s'aplica la mesura *tf*idf* per determinar les paraules més freqüents[4], s'incorpora el procés de reconeixement de domini que hem mencionat en l'apartat anterior i s'ha desenvolupat un mòdul que determina la distribució espacial de la informació més important d'un text. Aquest darrer paràmetre es computa amb una mesura anomenada *Optimal Position Policy* (Lin i Hovy 1997): partint d'un corpus en què jutges humans han marcat les oracions que consideren més importants, s'apliquen tècniques d'aprenentatge automàtic per determinar quines són les posicions on es troba la informació més important en un text: el títol, el primer i darrer paràgraf, etc. El resum resultant s'avalua calculant la seva precisió i cobertura respecte de les paraules clau associades a cada un dels textos de què es fa el resum. Com hem apuntat abans, aquest mètode funciona especialment bé si es pot determinar a priori el gènere i el domini al qual pertany el document que es tracta, tot i que també s'han detectat patrons generals de distribució espacial de la informació en textos sense restriccions. Així, per exemple, es pot determinar que la primera oració del primer i darrer paràgraf d'un text no literari qualsevol conté, molt probablement, part de la informació més rellevant de tot el text.

Sembla que la principal limitació dels mètodes que exploten les característiques documentals dels textos ve donada pel fet que deixen de banda la seva dimensió lingüística. És per això que la representació que s'obté dels textos és parcial, ja que tan sols es té en compte una part de la seva informació estructural. Normalment, es determina a priori quina ha de ser l'organització de la informació en un text donat, a partir d'un cert coneixement sobre el format dels diversos tipus de documents. D'aquesta manera, no queden recollides possibles particularitats del text, excepcions de la norma establerta que, sovint, solen ser molt informatives del seu contingut. No obstant, també és cert que algunes d'aquestes característiques documentals poden generalitzar-se per la major part dels textos: per exemple, el títol i el primer paràgraf solen contenir informació molt important en els textos de qualsevol gènere, i gairebé sempre paraules clau com 'per exemple' expliciten la manca d'importància d'un segment de text. És possible, doncs, que algunes característiques documentals es donin de manera prou general en els textos com per a poder ser considerades com a pròpiament lingüístiques. Per aquesta via, podem arribar a establir un vincle entre la dimensió documental i la natura lingüística dels textos, una conjunció del coneixement lingüístic i documental per a un objectiu comú: el resum automàtic.

Una altra aproximació al resum automàtic és la que explota les marques d'estructura de natura lingüística. En aquesta línia, s'intenta donar una caracterització completa dels textos mitjançant els anomenats '**marcadors del discurs**', marques formals superficials de l'organització dels textos, paraules clau amb un contingut estructural. Cal distingir entre aquestes i les paraules clau de contingut semàntic, que hem mencionat en l'apartat anterior. En ambdós casos es tracta de paraules que caracteritzen els textos, però els marcadors del discurs donen compte de la seva estructura, mentre que els 'marcadors lingüístics' (Lehmam 1995) en recullen el contingut. No obstant, hem de dir que la frontera entre aquests dos tipus de marcadors és molt difusa, i que un concepte ampli de 'marcador del discurs' pot incloure unitats que estan molt lluny de les marques

[4]Com hem dit en l'apartat anterior, aquests investigadors també han treballat sobre la freqüència de conceptes, però aquest mètode encara no ha passat de ser un experiment de laboratori.

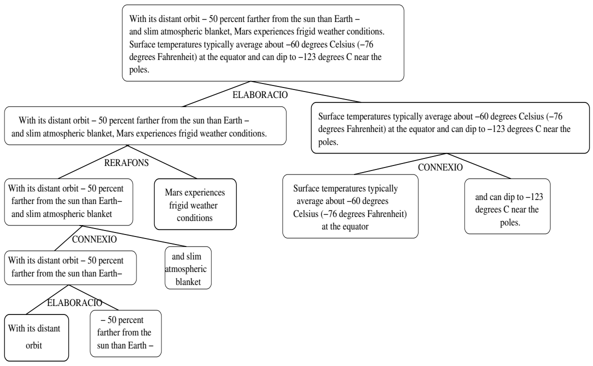

Figura 2.1: Representació arbòria de l'estructura retòrica d'un text (Marcu 1997: 18)

lèxiques invariables que es proposen com a marcadors prototípics. Veiem el següent exemple, que cita Bosque (2001):

> `Si uno` ***empieza por*** `permitirse un asesinato,` ***pronto*** `no le da importancia al robo, del robo` ***pasa a*** `la bebida y a la inobservancia del día del Señor, y` ***acaba por*** `faltar a la buena educación` ***y por*** `dejar las cosas para el día siguiente.`
>
> Thomas D. Quincey
> *Del Asesinato Considerado Como Una De Las Bellas Artes*
>
> En aquest exemple, l'efecte còmic ve donat perquè s'estableix una gradació entre els elements que és precisament la contrària a la que esperaria el lector. Es dóna, doncs, un contrast entre les inferències a què porten les paraules de contingut i les paraules d'estructura (en cursiva). Aquestes paraules d'estructura, però, no s'adiuen amb el concepte prototípic de marcador del discurs, sinó que pertanyen a classes obertes, tenen un important contingut lèxic i tenen usos no discursius habituals.

Ono et al. (1994) i Marcu (1997) van més enllà del concepte tradicional de 'paraula clau' que proposa Edmunson per explotar un conjunt molt més ampli de partícules discursives que aporten informació sobre certa estructura dels textos. Totes dues aproximacions parteixen de la teoria del discurs que proposen Mann i Thompson (1988) per obtenir una representació de l'estructura argumentativa del text, l'anomenada **estructura retòrica**. Aquesta estructura és una representació jeràrquica, amb forma arbòria, en què ha de quedar reflectida l'organització de la informació del text. L'arrel de l'arbre representa el nucli argumentatiu del text, i les diverses parts del text se situen en les branques de l'arbre d'acord amb les relacions que estableixen amb aquest nucli o entre elles. Així, com més directament relacionat estigui un fragment de text amb el nucli discursiu, més proper de l'arrel es trobarà el node que el representa en l'estructura arbòria. Inversament, aquelles parts del text que estableixen una relació indirecta amb el nucli discursiu, com les que aporten informació sobre parts menys centrals del text, estaran molt més allunyades de l'arrel, ja que la seva relació amb el nucli estarà mediada pels nodes que representen les parts del text amb què estableixen relacions directes. Al seu torn, aquests nodes poden estar relacionats directament amb el nucli discursiu o amb altres nodes. D'aquesta manera, com menys directa sigui la relació d'una part del text amb el nucli oracional, més allunyat de l'arrel de l'arbre estarà el node que el representa. Així doncs, la rellevància discursiva d'una part del text queda determinada per la seva posició relativa en l'estructura arbòria. (vegeu figura 2.1)

Per obtenir la representació de l'estructura retòrica dels textos, els investigadors parteixen de dues fonts de coneixement lingüístic: l'aparell teòric de la Rhetorical Structure Theory (RST, Mann i Thompson 1988) i els estudis sobre els marcadors del discurs de la llengua que tracten, el japonès en el cas d'Ono, Sumita i Miike i l'anglès en el cas de Marcu. A més, Ono et al. també utilitzen informació sintàctica parcial sobre els textos. Partint d'aquesta informació lingüística de base, aquests investigadors desenvolupen dos sistemes que intenten realitzar una anàlisi retòrica automàtica dels textos. Com a resultat, obtenen una representació jeràrquica del text que explicita la rellevància relativa de cada part del text. Per elaborar un resum automàtic a partir d'aquesta representació, s'han desenvolupat un conjunt d'heurístiques que eliminen progressivament les parts més accessòries d'un text, és a dir, les que estan situades en els nodes menys centrals de l'estructura arbòria. Així, el volum del text es va reduint fins a obtenir un text molt més curt que l'original però que recull els fragments discursivament més rellevants. D'aquesta manera es garanteix l'adequació del resum als continguts més importants del text i, alhora, es conserva la coherència, ja que es recull la línia argumentativa original.

La representació dels textos que es pot obtenir a partir de l'anàlisi de la seva estructura retòrica és certament parcial, ja que deixa de banda altres tipus d'informació també apta per a la caracterització dels textos. No obstant, si comparem aquesta aproximació amb un enfocament purament documental, veiem que la primera és més adequada a la realitat dels textos. Per una banda, l'anàlisi retòrica no tracta d'aplicar una estructura informació determinada a priori, sinó que obté la informació a partir de les marques que es troben efectivament en els textos. D'aquesta manera tenim una certa garantia que no es passarà per alt informació rellevant, ja que s'identifiquen les marques estructurals que caracteritzen les parts més importants d'un text. Per altra banda, l'anàlisi retòrica és aplicable a tot tipus de textos, sense les restriccions de gènere o domini pròpies de les aproximacions documentals. Finalment, la representació d'un text que podem obtenir a partir d'una anàlisi retòrica pot arribar a una profunditat que va molt més enllà de les possibilitats dels mètodes documentals i que s'acosta, en canvi, a la profunditat que busquen les tècniques d'Extracció d'Informació.

La principal limitació dels sistemes que parteixen de la detecció i processament de marques superficials de l'estructura dels textos, ja sigui la posició de les paraules en un text o bé la presència de certes paraules clau, és precisament la seva forta dependència d'aquestes marques. Les relacions textuals no es realitzen tan sols mitjançant peces lèxiques amb una funció discursiva fossilitzada, sinó que es poden vehicular a través de múltiples recursos lingüístics: les propietats semàntiques de les paraules, els mecanismes de focalització, els patrons de pronominalització i d'elisió, els temps i aspectes verbals, els tipus d'oració, etc. La lingüística textual i l'anàlisi del discurs han tractat amb profunditat aquest tipus de fenòmens, amb l'objectiu de descriure els textos de manera adequada, donant compte, sistematitzant i explicant aquelles característiques que els constitueixen com a objectes lingüístics diferenciats d'altres tipus d'informació i també d'altres tipus d'objectes lingüístics. Desenvoluparem alguns treballs que s'han dut a terme en aquesta línia en l'apartat 2.2.

Jing (1998) intenta detectar, sistematitzar i explotar marques estructurals d'un espectre més ampli que els marcadors del discurs. En el seu treball, Jing parteix de la hipòtesi que algunes de les estratègies de què disposen els humans per elaborar un resum d'un text consisteix a **adaptar** fragments literals del text original perquè formin part del resum. La seva proposta consisteix a identificar aquestes estratègies, per tal de desenvolupar un sistema de resum que les reprodueixi. En un estudi en què contrasta textos i resums realitzats per professionals de la documentació, Jing troba les següents estratègies d'adaptació de fragments:

- **eliminació** de components oracionals accessoris.
- **combinació** d'oracions per eliminar informació redundant, com per exemple un subjecte repetit.
- **transformacions** sintàctiques o **paràfrasis** lèxiques (generalitzacions, especificacions) que redueixin el volum del text original o contribueixin a la seva comprensió.

– **reordenació** de fragments de manera que es redueixi el volum del text perquè es puguin realitzar operacions de combinació d'oracions o perquè la reordenació contribueixi a la millora de la comprensió.

En la primera de les estratègies que exposem, Jing explota marques superficials de certa estructura del discurs, a partir de les quals pot determinar de manera automàtica la rellevància o manca de rellevància de segments inferiors a l'oració. Aquestes marques estan incloses en el que podríem considerar marcadors del discurs, però Jing també té en compte informació morfosintàctica (oracions de relatiu, sintagmes adverbials...). Per explotar de manera efectiva aquesta informació, Jing l'associa a certs patrons de puntuació o d'ordre de components de les oracions que evidencien i fan inambigua la funció discursiva d'aquests elements. Per exemple, una frase circumstancial en posició final d'oració i precedida de coma serà susceptible de ser sotmesa al procés d'eliminació.

La diferència més evident amb les propostes basades en una teoria sobre el discurs es troben en el fet que Jing no obté, ni pretén obtenir, una representació global del text. El seu sistema utilitza altres tècniques (cadenes lèxiques, posició en el document) per identificar les parts més rellevants del text. Un cop determinats quins són els fragments més rellevants, s'els aplica les estratègies d'adaptació de fragments que hem exposat, i obté un text final més reduït. A més, Jing tampoc no considera el tipus de relació que poden establir els segments de text relacionats per una marca d'estructura, sinó que tan sols determina si un dels segments pot ser eliminat en un resum. La proposta de Jing, doncs, soluciona una de les principals limitacions dels sistemes d'Ono et al. i Marcu, la seva forta dependència dels marcadors del discurs. No obstant, hem de dir que el sistema de Jing no pot obtenir la profunditat descriptiva a què poden arribar les aproximacions basades en l'anàlisi d'estructura retòrica.

També apunta en el treball d'Edmunson una tendència que es generalitzarà més tard en el camp del resum automàtic: la **integració** de diverses mesures de rellevància que, de manera col·laborativa, contribueixen a donar una representació més completa d'un text. Certament, els textos són un conglomerat d'informacions diverses, per tant, si volem descriure'ls de manera adequada, sembla clar que haurem de donar compte de les seves facetes diverses. Així, els sistemes que funcionen en l'actualitat tendeixen a incorporar diversos tipus d'informació per determinar quines són els elements més importants d'un text, que hauran de ser reflectides en un resum. Ara bé, creiem que no està encara prou estudiat com aquests tipus diversos d'informació sobre els textos interactuen, i com contribueixen de manera conjunta a determinar la rellevància en els textos. Molt possiblement, si aconseguim d'aprofundir en aquesta qüestió, es produirà un salt qualitatiu en el resum automàtic, ja que es podran tenir en compte els textos en les seves diverses facetes de manera integrada: no només comptarem amb més informació sobre els textos, sinó que la informació de què disposarem serà més útil.

Una bona mostra de com diverses mesures poden contribuir a la millora del resum final és el sistema de Kupiec et al. (1995), en què col·laboren informacions tan diverses com la longitud de les oracions, la posició de les oracions en el paràgraf, paraules clau de contingut, paraules en majúscules, etc. A més, utilitzen tècniques d'aprenentatge automàtic per determinar el valor relatiu de cadascuna d'aquestes informacions i les possibles interaccions que s'hi donin. Els resums obtinguts, amb un 20% del volum inicial, donen un 74% de millora respecte d'un resum que prengui el primer 10% del text original.

2.1.3 Comprensió Profunda

Una tercera via per a l'elaboració d'un resum automàtic és la que intenta dur a terme una comprensió profunda dels textos. En general, aquesta comprensió es realitza tractant de detectar la informació més important d'un document a través de cert coneixement del món aplicat de manera

The afternoon of February 26, 1993, Reuters reported that a suspected bomb killed at least five people in the World Trade Center. Later the same day, Reuters announced that exactly five people were killed in the blast.

MESSAGE: ID	TST3–MUC4–0010
INCIDENT: DATE	01 NOV 89
INCIDENT: LOCATION	EL SALVADOR
INCIDENT: TYPE	ATTACK
INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
INCIDENT: INSTRUMENT TYPE	–
PERP: INCIDENT CATEGORY	TERRORIST ACT
PERP: INDIVIDUAL ID	"TERRORIST"
PERP: ORGANIZATION ID	"THE FMLN"
PERP: ORG. CONFIDENCE	REPORTED: "THE FMLN"
HUM TGT: TYPE	CIVILIAN: "1 CIVILIAN"
HUM TGT: NUMBER	1: "1 CIVILIAN"
HUM TGT: EFFECT OF INCIDENT	DEATH: "1 CIVILIAN"

Figura 2.2: Extracte d'una templeta del concurs MUC-4 (Radev 1996)

deductiva sobre el text. La línia més desenvolupada és la que intenta trobar en els textos la informació especificada a priori en un esquema que caracteritza els trets més rellevants d'un cert tipus de textos. Aquests esquemes, també anomenats 'scripts' o 'templetes' (vegeu figura 2.2), són elaborats a partir d'una modelització del coneixement requerit per a la comprensió d'aquests textos en un procés computacional. Aplicant aquestes estructures de coneixement sobre documents del tipus determinat, s'obté una estructura del text en què es recull la informació considerada més important.

Aquesta mena d'enfocaments ha estat àmpliament treballada en extracció d'informació i en generació de llenguatge, dues àrees del PLN que requereixen una comprensió profunda del llenguatge. Intuïtivament, sembla que l'elaboració d'un resum també ha de passar per una comprensió del text a resumir, per tant, una aproximació d'aquest tipus resultaria també molt adequada per a tasques de resum automàtic. No obstant, trobem relativament pocs treballs que desenvolupin aquesta línia, per dues raons bàsiques: el seu alt cost de desenvolupament i el fet que es poden aplicar a tipus de textos molt restringits per gènere documental o domini conceptual.

Schank i Abelson (1977) presenten una primera aproximació al resum automàtic amb comprensió profunda, amb una aplicació del seu sistema d'EI basat en scripts, modelitzacions d'algunes situacions de la vida quotidiana en què s'identifiquen els principals events i s'intenta reconèixer les intencions dels participants. Però el primer exponent destacat d'aquest tipus d'aproximació és el sistema FRUMP (DeJong 1979). Aquest sistema tracta els teletips d'agències de notícies en dues fases: en primer lloc, busca evidència que una notícia pertany a alguna de les classes d'events predefinits, i, si és així, hi aplica l'esquema corresponent. En aquest esquema s'estipulen les dades considerades més importants en un event d'un tipus determinat. Aquestes dades es localitzen en el text, s'extreuen i s'utilitzen per l'elaboració d'un text més curt. D'aquesta manera, el resum ha de recollir la informació més rellevant, perquè es considera la més caracteritzadora, d'una notícia. Ara bé: com que els esquemes que estipulen la informació a extreure s'elaboren a priori, és possible que no reflecteixin la realitat dels textos de manera adequada, ja que no es podrà donar compte de la informació que no es reculli en els paràmetres preestablerts en la caracterització de cada tipus de notícia. I encara més: si no es troba cap estructura de dades adient a aquella notícia, la notícia no podrà ser tractada en absolut.

Malgrat les evidents limitacions d'una aproximació de tipus deductiu, el sistema MultiGen (Radev i McKeown 1998) mostra com les possibilitats d'un sistema que es basi en coneixement profund van molt més enllà d'allò que permeten els mètodes superficials. El sistema MultiGen parteix de les templetes del concurs MUC, en què concorren diversos sistemes d'EI, i gestiona la informació que proporcionen per elaborar el resum d'un event. En primer lloc, s'agrupen les notícies que tracten un mateix event. Seguidament s'identifiquen els conceptes característics de l'event i es comparen els atributs associats a cadascun d'aquests conceptes en cada notícia. D'aquesta manera es determinen les característiques definitòries de l'event en qüestió i es dóna compte de la informació

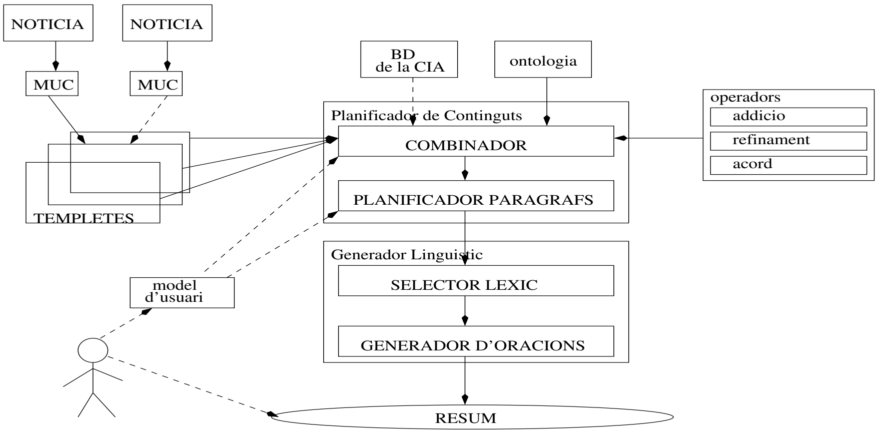

Figura 2.3: Arquitectura del sistema MultiGen

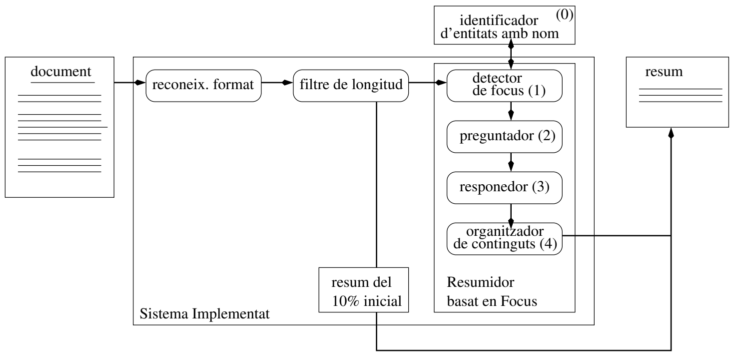

Figura 2.4: Arquitectura del sistema FociSum

associada als fets i entitats que hi ocorren, però, a més, es pot donar compte de les discrepàncies entre notícies de diferents fonts (agències de notícies) pel que fa a la informació sobre l'event. Per combinar les templetes de cada notícia de les diverses fonts utilitzen operadors lògics que identifiquen els principals modes de discrepància: canvi de perspectiva, contradicció, refinament i acord, generalització, etc. (vegeu figura 2.3)

Entre els avantatges que presenta aquest sistema, podem destacar la realització de resums de múltiples documents, és a dir, de diverses notícies sobre un mateix fet. MultiGen també sintetitza la informació redundant i dóna compte de les discrepàncies entre diverses fonts d'informació. A més, transforma l'estructura d'informació d'una templeta MUC en llenguatge mitjançant un mòdul de generació de llenguatge natural, cosa que garanteix una coherència major que en els sistemes basats en extractes. Ara bé, la profunditat que ofereix aquest sistema té com a contrapartida un camp d'aplicació molt restringit, ja que tan sols pot aplicar-se a teletips de notícies sobre atemptats terroristes.

Les fortes restriccions de domini que presenta el sistema MultiGen van portar el mateix equip d'investigadors a desenvolupar el sistema FociSum (Kan i McKeown 1999). Aquest sistema aplica l'estratègia dels esquemes d'extracció d'informació a petita escala, de manera que no intenta donar compte del conjunt del text amb un sol esquema, sinó que centra els esforços en les parts considerades més importants, que han estat prèviament identificades per un procés de detecció d'entitats (vegeu figura 2.4). D'aquesta manera, es realitza la comprensió profunda dels fragments de text que realment formaran part del resum, mentre que s'estalvien esforços de processament de parts poc rellevants del text. De la mateixa manera, aquesta estratègia introdueix un punt de vista més inductiu, que dóna una certa garantia que no es passaran per alt elements informatius no contemplats en la modelització prèvia del món: s'adapta la comprensió del text al text mateix, i no a la inversa. Com a conseqüència, l'anàlisi resultant del text dóna compte de la informació que contenen certs fragments, i no de la totalitat del text, tot i que en alguns casos també es poden detectar les relacions entre els events i les entitats identificats en aquests fragments. Com a principal avantatge, hem de dir que el sistema FociSum és el més transportable dels sistemes d'extracció d'informació basats en templetes que coneixem.

Sembla evident que, com més profunda sigui l'anàlisi del text, més adequada serà la caracterització que se n'obtingui i millor serà el resum final que se'n pugui elaborar. No obstant, hem de tenir en compte que els mètodes per a l'anàlisi profunda del llenguatge encara no han assolit un desenvolupament òptim per a dominis irrestrictes, de manera que els resultats que ofereixen no són totalment fiables. A més, hem de tenir en compte que aquests mètodes s'han treballat sobretot en l'àmbit anglòfon i, sovint, per a gèneres documentals tancats. És per això que, malgrat que ofereixen possibilitats d'anàlisi del llenguatge que no hem d'oblidar, les limitacions actuals del PLN per llengües diferents de l'anglès fa que aquest tipus d'aproximació resti encara inabastable en la major part dels casos.

2.1.4 Conclusions

Després d'aquesta revisió, podem arribar a dues conclusions sobre l'evolució dels mètodes de resum automàtic. En primer lloc, constatem una tendència al desenvolupament i aprofundiment de les tècniques superficials de tractament dels textos. Pel que fa a la freqüència de paraules, els documents ja no es processen en termes absoluts, sinó que s'ha incorporat la qüestió contextual, és a dir, que un document sol formar part d'una col·lecció de documents, encara més, d'un conjunt de documents caracteritzat per paràmetres comuns, i, alhora, que es dóna en una llengua determinada: aquests factors contribueixen a matisar el valor de la freqüència de paraules en un document, amb mesures relatives com la *tf*idf* i la incorporació de cert coneixement lingüístic sobre el lèxic, com són les llistes de paraules buides. Respecte dels mètodes que utilitzen paraules clau, sembla que la tendència és a l'ampliació del seu abast: s'identifiquen cada cop més elements susceptibles d'aportar informació com a paraules clau, i la informació associada a aquests elements és cada cop més profunda: així ho mostren, per exemple, els treballs que exploten el format de gènere o els que intenten sistematitzar les marques d'estructura del discurs.

Una segona conclusió, potser més agosarada, constata la tendència oposada: trobem que alguns sistemes de resum automàtic que proposaven una comprensió molt profunda dels textos han reduït la profunditat de la seva proposta. És el cas dels sistemes que treballen a partir de templetes: per exemple, a partir del sistema MultiGen, que treballava amb una única templeta global de tot un text, s'ha desenvolupat un sistema de templetes dinàmiques, més petites i adaptables a unitats discursives menors que el text, com ara les entitats del discurs. En altres casos, trobem que alguns investigadors han optat per aplicar un model teòric que es basa en la comprensió profunda dels textos de manera superficial: seria el cas dels qui apliquen la RST. D'aquesta manera, pretenen assolir una major independència de gènere o domini. Malgrat que això suposi una certa pèrdua de profunditat en el processament des textos, podem considerar que es tracta d'una abstracció de les aproximacions més profundes, una generalització sobre els fets del llenguatge que, a mig o llarg termini, pot portar a aproximacions més globals, i no per això menys profundes, al tractament dels textos.

Sembla, doncs, que els esforços en resum automàtic resulten especialment efectius per als mètodes que pretenen assolir una certa comprensió del text, i, per tant, que incorporen un cert coneixement lingüístic o documental, però que mantenen una superficialitat i robustesa que els fa aptes per a les aplicacions informàtiques que puguin respondre a les necessitats de la societat de la informació en què vivim. Entre les aproximacions d'aquestes característiques, trobem especialment interessants les propostes que tracten el nivell discursiu dels textos, ja que aborden un aspecte lingüístic poc tractat en el PLN i en la LC però que pot resultar molt informatiu i, sens dubte, molt interessant.

La línia de RA que aborda el tractament de l'estructura discursiva dels textos sembla força adequada a la situació actual del PLN espanyol. Si explotem les marques superficials d'estructura podem adequar-nos a aquestes limitacions, i alhora obtenir una representació adequada dels textos. Si bé és cert que la informació discursiva que trobem en els textos és molt heterogènia, hem cregut que, en un nivell d'anàlisi superficial, la informació discursiva més útil per a resum automàtic seria aquella que dóna compte de l'organització de les idees en els textos, tal com proposen Ono, Miike i Sumita (1994) i Marcu (1997). No obstant, cal tenir molt en compte la proposta de Jing (1998), ja que soluciona algunes de les limitacions de l'aproximació retòrica.

Tot seguit presentem l'estat de la qüestió en l'estudi computacional de l'estructura dels textos, centrant-nos especialment en aquells que responen més adequadament als nostres objectius d'investigació.

2.2 El discurs en el PLN

D'ençà de l'adveniment de la pragmàtica i la lingüística del text, s'han dedicat molts esforços a donar compte de l'organització del discurs. Tal com exposen Dressler i Beaugrande en la seva introducció a la lingüística del text (Dressler i Beaugrande 1981), es donen diversos factors que constitueixen els textos com a objectes lingüístics cohesius, coherents i relacionats amb la situació comunicativa en què es donen, alhora que diferenciats de les unitats d'altres nivells lingüístics: fenòmens de co-referència (repetició, paràfrasis, anàfora, elisió), de distribució de la informació (nou - conegut), de coherència semàntica i pragmàtica, de desenvolupament de l'argumentació, i també les relacions amb l'entorn comunicatiu: amb les intencions comunicatives i els estats atencionals dels participants, amb la resta de textos que comparteixen l'espai comunicatiu possible, amb el coneixement del món que comparteixen els parlants, etc.

Malgrat que tots aquests aspectes contribueixen de manera conjunta a la constitució del text, tractar-los tots alhora suposa una anàlisi profunda i extensa que requereix un gran esforç. És per això que els diversos treballs sobre el text i el discurs han tendit a explorar cadascun d'aquests aspectes de manera aïllada, per arribar a donar compte de manera adequada, encara que parcial, de les característiques més rellevants o més reveladores dels textos. Això és especialment evident quan es tracta d'aproximacions computacionals als textos, ja que l'objectiu aplicat de la major part d'aquestes aproximacions suposa una limitació més al poder explicatiu de les teories sobre el discurs. Tot seguit presentem algunes propostes del tractament computacional de l'estructura discursiva dels textos, com a mostra, breu però significativa, de les possibles postures computacionals en aquest nivell lingüístic.

Una de les propostes més rellevants en la caracterització computacional del discurs és la teoria de Grosz i Sidner (1986). Proposen distingir tres nivells: el lingüístic, l'intencional i l'atencional. En el nivell lingüístic donen compte de l'estructura de la seqüència d'expressions que conformen el discurs, és a dir, divideixen el text en segments[5] i determinen a quin segment pertany cada part del text. En el nivell intencional, postulen que el discurs està dirigit per les intencions comunicatives dels participants. Així, donada una intenció general del discurs, es pot determinar per cada segment quina és la seva contribució a l'objectiu del discurs. Sovint, els objectius dels diversos segments individuals no contribueixen de manera directa a la intenció general del discurs, sinó que

[5]El concepte de segment amb què treballen Grosz i Sidner és molt allunyat del que utilitzem nosaltres en aquest treball, vegeu cap. 4

s'estableixen relacions entre els segments segons si l'objectiu d'un segment ha de ser satisfet abans que l'objeciu d'un altre, o bé si l'acompliment d'un segment del text és part de l'acompliment d'un altre. Aquestes relacions entre segments reben el nom de *satisfacció-precedència* i de *dominància*, respectivament (Grosz i Sidner 1986: 179). Aplicant recursivament aquestes relacions entre segments, es pot arribar a una estructura jeràrquica del text.

A partir de la identificació de les intencions que dirigeixen el discurs, Grosz i Sidner també intenten donar compte dels estats atencionals, és a dir: de quina informació és més rellevant en cada punt del discurs. Els estats atencionals es construeixen acumulant espais de focalització que contenen els objectes i relacions dels segments del discurs que es processen, fins que s'arriba a un canvi de focus, amb la qual cosa es canvia d'estat atencional i es construeix un conjunt diferent d'espais de focalització.

Pel que fa a la implementació computacional de la teoria de Grosz i Sidner, s'han treballat dues línies: per una banda, el desenvolupament de mètodes de comprensió profunda dels textos que aportin la informació necessària per al funcionament d'aquest model. Per altra banda, també s'han dut a terme treballs per trobar marques superficials que explicitin de manera sistemàtica les intencions comunicatives (Passonneau i Litman 1993) i també sobre les evidències de canvi de focus (Hirschberg et al. 1993), però sembla que cal realitzar una comprensió profunda del text per arribar a un nivell d'anàlisi acceptable pel que fa a les intencions comunicatives i al focus d'atenció.

Des del nostre punt de vista, la proposta de Polanyi (1988, 1995) i Scha i Polanyi (1988) és la perspectiva més integradora entre les que tracten el discurs des del punt de vista computacional. El Linguistic Discourse Model processa els textos des de la morfologia a les relacions intersegmentals, passant per la sintaxi de constituents i l'oracional, la semàntica lèxica i de proposicions, la progressió dels temps verbals, qüestions de punt de vista, etc. D'aquesta manera, assoleix una comprensió profunda del text que resulta especialment adequada perquè té en compte moltes de les diverses facetes que contribueixen a la constitució del text com a objecte lingüístic i comunicatiu. Ara bé: la proposta de Polanyi, malgrat el seu indubtable interès, està molt lluny de les nostres possibilitats d'aplicació, ja que actualment no disposem d'eines de PLN que ens donin una anàlisi profunda, satisfactòria i sense restriccions de domini en cadascun dels nivells lingüístics que hem enumerat abans. És per això que, malgrat que no descartem arribar al nivell de comprensió del discurs que descriu aquesta investigadora, ara com ara hem d'atenir-nos a abordatges més reduïts o més superficials.

Per al nostre propòsit, el resum automàtic, i les característiques dels recursos de PLN, ens semblen especialment adequades aquelles perspectives que tracten les anomenades 'relacions de coherència', és a dir, la manera com diversos segments de text s'organitzen entre ells en el desenvolupament d'un text. Tot seguit esbossarem les aportacions, limitacions i dificultats de les principals teories computacionals que tracten les relacions de coherència en el discurs.

2.2.1 La coherència textual

Les qüestions de coherència textual han estat especialment tractades des de l'àrea de generació del llenguatge natural. Això és així perquè en la generació de llenguatge es requereix un coneixement extens i profund juntament amb mecanismes d'organització d'aquest coneixement, que solen basar-se en models de cognició humana. Les relacions de coherència, per la seva plausibilitat cognitiva, semblen uns mecanismes òptims per l'organització del coneixement requerit per al processament profund dels textos. Ara bé, com hem dit abans, també es requereix un coneixement profund i extens en altres aplicacions de PLN, sobretot les que tracten amb una certa profunditat els nivells lingüístics més complexos, com el text. En aquesta darrera categoria trobaríem el resum automàtic lingüísticament rigorós. Aquest tipus d'aproximació tracta de determinar el contingut d'un text, com s'estructura aquest contingut i com es distribueix el fluxe d'informació i la rellevància en aquesta estructura. L'objectiu és, doncs, arribar a la comprensió profunda del text per elaborar-ne un resum fidel.

Podem considerar Hobbs (1978, 1985) un dels pioners en l'estudi i sistematització del que es coneix com 'relacions de coherència' o 'relacions retòriques'. Potser precisament perquè Hobbs treballa en l'àrea de generació del llenguatge natural i de gestió de bases de coneixement, el seu estudi dels textos presenta una perspectiva marcadament cognitiva que contribueix a sistematitzar algunes de les qüestions més fonamentals pel que fa al tractament computacional d'aquest tipus d'informació. Per començar, explicita els supòsits que:

- els textos tenen estructura, és a dir, que no són una cadena de frases inconnexes
- els textos poden ser no atòmics, i els textos no atòmics es divideixen en segments
- donat que els segments d'un text es refereixen a conceptes del món relacionats entre ells, els segments estableixen relacions entre ells que reflecteixen aquestes relacions
- podem donar compte d'aquestes relacions en una estructura que hauria de reflectir certa estructura del text original

Partint d'aquí, proposa una tipologia de relacions de coherència entre segments, relacions que, en última instància, han de reflectir els processos cognitius de processament dels textos. La coherència de què intenta donar compte Hobbs està essencialment basada en els mecanismes d'inferència que possibiliten la construcció de la representació conceptual d'un text.

També Mann i Thompson (1988), en la seva Rhetorical Structure Theory (RST), treballen amb una sèrie de relacions formals, les anomenades 'relacions retòriques', per a obtenir un esquema jerarquitzat, en forma d'arbre, que representi l'estructura retòrica d'un text. Com en el cas de Hobbs, Mann i Thompson prenen com a referència un model cognitiu del parlant. La principal diferència entre aquests investigadors rau en el tipus de facultat lingüística en què es posa l'èmfasi: mentre Hobbs tracta de donar compte dels mecanismes de processament passius, és a dir, els que realitza el receptor d'un missatge per a comprendre'l, Mann i Thompson prenen un punt de vista de producció, parteixen de la intenció de l'emissor per explicar la línia argumentativa i d'organització de la informació d'un text. Tot i que ambdues aproximacions tracten de fer una teoria d'abast general sobre les relacions discursives en els textos, cal tenir en compte el tipus de produccions lingüístiques que estudia cada investigador: mentre que Hobbs intenta abastar una tipologia d'objectes lingüístics que va des del diàleg fins a les demostracions matemàtiques, Mann i Thompson treballen bàsicament sobre textos escrits de tipus expositiu i argumentatiu. Ara bé, sembla que ens trobem davant d'una teoria sobre l'organització del discurs prou general, ja que permet donar compte de tipus molt diversos de textos, i és prou flexible per incorporar característiques que no hagués tingut prèviament en compte.

La RST intenta donar les pautes generals per descriure les relacions entre segments de text, independentment de la presència de marques lèxiques o sintàctiques que les explicitin. La relació discursiva bàsica entre segments és la de nuclearitat: en la major part de relacions, un dels segments de text, el nucli, és més nuclear en la intenció comunicativa general i en el fluxe d'informació del text, mentre que un altre segment, el satèl·lit, aporta informació accessòria. A més, la relació entre nucli i satèl·lit és asimètrica, ja que el satèl·it depèn discursivament, sovint també sintàcticament, del nucli, però no a la inversa. Sobre aquest pilar bàsic, la RST intenta donar compte dels textos a partir dels següents elements:

Relacions relacions retòriques entre dos fragments de text no coincidents, dels quals un sol ser el nucli i un altre el satèl·lit. Es defineixen en quatre camps: restriccions en el nucli, restriccions en el satèl·lit, restriccions en la combinació nucli-satèl·lit i efecte de la relació en el lector. Mann i Thompson (1988) proposen una llista de relacions a tall d'exemple, però la deixen oberta a possibles ampliacions i modificacions.

Esquemes regles de combinació de relacions, força laxes, com també laxa sembla la sintaxi discursiva

Estructures (anàlisis estructurals) conjunt d'esquemes que acompleixen les condicions de

completitud: hi ha un esquema que conté tots els fragments de text de què consta el text

connexió: tots els fragments de text pertanyen a un esquema

unicitat: cada esquema conté un conjunt diferent de fragments de text

adjacència: els fragments de text de cada esquema constitueixen un fragment de text

Aquests elements constitueixen les eines amb què es realitza l'anàlisi dels textos que ha de resultar en una estructura que el representi per complet. Efectivament, en múltiples anàlisis dutes a terme per especialistes, sembla que és possible, per a la major part de textos, d'obtenir una representació que acompleixi els requisits de bona formació de les estructures de la RST.

La línia encetada per Hobbs i Mann i Thompson ha tingut una gran acceptació en el PLN, per la qual cosa trobem una multitud de treballs que tracten les relacions de coherència en els textos, sobretot des de l'àrea de generació del llenguatge natural. Moore i Pollack (1992) treballen en un dels principals punts febles de la RST: el fet que, malgrat que ho enuncia, no distingeix clarament entre relacions informacionals i relacions intencionals[6]. Creuen que cal donar compte de les relacions entre segments de text en dos nivells, en què es distingeixin clarament aquests dos tipus d'informació, perquè si no, la representació final del text resultara incompleta, si no es té en compte un dels dos nivells, o inadequada, si es barregen informacions de tots dos.

Resulta especialment representatiu el treball de Hovy i Maier (1997), en què recullen i sistematitzen un nombre important de relacions de coherència textual que es troben en la bibliografia fins al moment i intenten integrar-les en una tipologia única. En aquesta tipologia es mantenen els dos grans grups de relacions, informacionals i textuals, però sorgeix un tercer tipus d'informació relacional, clarament motivat per la seva utilitat en la generació de llenguatge natural: es tracta de les relacions textuals, que inclouen conjuncions lògiques (conjunció, disjunció), seqüències i llistes.

2.2.2 Coherència en un nivell superficial

La major part de treballs en la línia de les relacions de coherència (Hobbs 1985, Mann i Thompson 1988, Moore i Pollack 1992, Hovy i Maier 1997) parteixen del supòsit que, per detectar o generar aquestes relacions en els textos, es disposa d'una base de coneixement del món per donar compte del significat dels segments d'un text i dels mecanismes d'inferència necessaris per al processament del discurs. Donat que nosaltres no disposem d'aquesta base, aquestes propostes constitueixen una referència teòrica a tenir en compte a l'hora de definir i delimitar el tipus d'informació textual de què volem donar compte, però hem d'adoptar una aproximació més superficial al seu tractament computacional.

Creiem que les relacions retòriques evidencien els mecanismes de comprensió textual, els quals, al seu torn, són una mostra particular de mecanismes cognitius generals molt rendibles i, per tant, molt utilitzats. És per això que en trobem marques explícites en els textos: segons algunes teories cognitives, els mecanismes més utilitzats són els més marcats, perquè això suposa una major economia de processament (Knott i Dale 1994). I són precisament aquestes marques explícites i

[6]En terminologia de la RST, *subject-matter* (informacionals) són aquelles relacions que tan sols han de ser reconegudes pel lector, mentre que les *presentational* (intencionals) volen provocar una inclinació en el lector.

sistemàtiques de les relacions retòriques que fan possible d'extreure automàticament informació sobre l'organització de les idees en els textos. Ara bé, aquestes marques es poden donar en diversos nivells lingüístics: en les partícules de connexió discursiva, en els patrons de pronominalització, en l'ordre de paraules, en la distribució de temps verbals, en la semàntica proposicional, etc. Algunes d'aquestes relacions requereixen una comprensió profunda dels textos, que sovint implica un important coneixement afegit (anàlisi morfosintàctica, semàntica i discursiva dels textos) i un alt cost de processament, com en el model de Polanyi (1988, 1995). En canvi, les relacions expressades per peces lèxiques amb un funcionament discursiu sistemàtic i per signes de puntuació són detectables mitjançant una anàlisi superficial del text, amb procediments computacionals poc costosos com el reconeixement de patrons.

Ara bé: no sempre cal o és adequat de dur a terme una anàlisi profunda dels textos. Per exemple, si prenem una postura excessivament semanticista i pretenem donar compte del discurs a través exclusivament de la semàntica proposicional i les inferències que es poden realitzar a partir d'aquesta informació i del coneixement del món que se suposa que té un parlant, certa informació discursiva no queda recollida. Vegem els següent exemple de Montolío (2001) (pg. 19 i 20):

exemple: *"Pronto hará buen tiempo" - "Iré al gimnasio cada tarde"*

En aquest exemple, qualsevol parlant competent de l'espanyol entén sense problemes el contingut proposicional de cadascuna de les oracions, però no té manera de saber quina relació lògico-semàntica s'estableix entre elles. Si aplica el seu coneixement del món, probablement arribarà a les següents hipòtesis de relació entre els enunciats:

condició "Si *hace pronto buen tiempo, iré al gimnasio cada tarde"*

causa *"***Como** *pronto llegará el buen tiempo, iré al gimnasio cada tarde"*

contraargument *"***Aunque** *pronto hará buen tiempo, iré al gimnasio cada tarde"*

Tan sols amb la presència explícita de les marques discursives podem saber quina de les tres opcions més plausibles és la relació que l'emissor ha volgut reflectir efectivament.

Sembla, doncs, que una anàlisi superficial del discurs pot donar compte de certs fenòmens que no quedarien recollits per certes anàlisis més profundes dels textos. És per això que no podem considerar una anàlisi superficial com a poc informativa: la informació que podem obtenir d'una anàlisi és qualitativament diferent de la informació resultant de processos que treballen amb més coneixement lingüístic, però trobem que aquests tipus diversos d'informació es complementen i s'enriqueixen, i que cal tenir en compte les diverses facetes dels textos per arribar a una representació completa i satisfactòria.

Una proposta interessant per al tractament superficial de les relacions textuals és la que presenten Ono et al. (1994) i Marcu (1997): a partir del formalisme de la RST, intenten detectar i sistematitzar la informació retòrica associada als anomenats 'marcadors dels discurs', peces lèxiques o sintàctiques que tenen un significat i funció específicament discursius. Aquesta aproximació resulta interessant perquè, donades les limitacions actuals del PLN, possibilita la implementació efectiva de sistemes de tractament computacional del discurs, de manera que es pot treballar un nivell lingüístic molt complex de manera superficial. Creiem que la superficialitat d'aquesta aproximació no està renyida amb la profunditat dels fenòmens que tracta, ans al contrari: basant-se en una teoria prou ambiciosa del discurs, tracta fenòmens altrament inabordables. Les mancances derivades de la seva superficialitat repercuteixen, més aviat, en la profunditat amb què es tracten aquests fenòmens, però no en la profunditat dels fenòmens en ells mateixos.

L'anàlisi retòrica dels textos proporciona independència de gènere i domini perquè l'únic requisit necessari per a la seva aplicació és la bona formació discursiva del text a analitzar (Marcu 1997, Montolío 2001), de la mateixa manera que el requisit fonamental per a l'anàlisi sintàctica de text lliure és la bona formació gramatical de l'oració. La bona formació és una característica inherent als objectes lingüístics, i un parlant competent d'una llengua no hauria de realitzar textos o oracions mal formades, com a mínim en un registre escrit i formal. Tanmateix, en la realitat dels textos podem trobar seqüències mal formades. En aquest cas, no podrà realitzar-se una anàlisi retòrica satisfactòria del text, perquè no es contempla la mala formació. No obstant, els sistemes de PLN esdevenen cada cop més robustos, amb l'objectiu final de poder tractar, de la mateixa manera que ho fem habitualment els humans, objectes lingüístics que violen les regles de bona formació.

exemple: *Actualmente los estudiantes no pueden acceder a la carrera que ambicionan y,* ***pese a todo****, la masificación es un hecho en las aulas universitarias;* ***pues****, la solución más democrática es aumentar las restricciones académicas y* ***a no ser que*** *aumenten las restricciones económicas.*

Aquest exemple de Montolío (2000: 120) posa de manifest com la mala formació discursiva dels textos dificulta la seva interpretació, automàtica però també humana. En aquest cas concret, la principal font d'incoherència es troba en els marcadors del discurs, per la qual cosa una anàlisi basada en coneixement associat a aquests elements seria totalemtn errònia.

Hem dit, doncs, que l'anàlisi retòrica dels textos es basa en la bona formació discursiva. Si establim un paral·lelisme entre la noció de bona formació discursiva i la bona formació gramatical, veurem que totes dues tracten de reflectir la competència lingüística dels parlants, però en nivells lingüístics diferents. La diferència més important rau en la formalització de cadascun d'aquests conceptes[7]. En el nivell oracional disposem conjunts de regles que donen compte de com es construeixen oracions i expliquen per què una determinada seqüència de paraules no constitueix una oració correcta. En canvi, en el nivell discursiu no disposem d'un aparell formal de referència que expliqui com s'organitzen els diversos components del discurs per a formar textos lingüísticament acceptables. Davant d'aquesta mancança, recorrem a la noció de 'felicitat' discursiva per donar compte de les seqüències textuals que un parlant competent de la llengua acceptaria. Aquesta noció seria l'equivalent discursiu de la gramaticalitat, però presenta una vaguetat molt més acusada.

Ara bé: per dur a terme una anàlisi automàtica de certa estructura discursiva dels textos no podem basar-nos en vagues nocions que apel·len a les intuïcions dels parlants. Es fan necessàries marques formals que explicitin les relacions de coherència textual, és a dir, s'ha de poder donar compte d'aquesta coherència de manera prou formal per poder processar el text amb mitjans computacionals. Trobem que els textos escrits de registre formal i amb una longitud considerable realitzen de manera més explícita, en un nivell lingüístic més superficial, els fenòmens de coherència textual. Això ve donat pel fet que en aquests contextos esdevé necessari guiar les inferències dels lectors. En canvi, en altres contextos comunicatius no trobem tanta abundància de marcadors que explicitin els mecanismes necessaris per a la comprensió del text, perquè la proximitat dels referents (en el cas de les situacions de conversa) o la limitació del camp de referència possible (en el cas del discurs de gènere) fan que les inferències resultin molt menys costoses. Tanmateix, en aquestes altres situacions comunicatives trobem una gran abundància de marcadors del discurs, però amb funcions ben diferents: la regulació dels torns de parla, la clarificació de la relació entre els participants, etc.

Malgrat els avantatges que presenta l'anàlisi retòrica dels textos per sobre d'altres mètodes d'anàlisi discursiva, no és suficient per obtenir una representació completa i adequada del text original, que solucioni de manera satisfactòria la generació automàtica de resums. Com indiquen Ono et al. (1994), els algorismes d'anàlisi retòrica actuals resulten efectius tan sols en textos amb gran

[7]Sempre per a la llengua espanyola.

profusió de marcadors discursius explícits, però podem trobar textos que basin la seva coherència en les propietats semàntiques de les paraules i proposicions que constitueixen el propi text, prescindint de cap aportació de marcadors del discurs. Tanmateix, alguns estudis quantitatius (Redeker 1990, presentem les nostres pròpies dades en l'apartat 3.2.1) mostren que es dóna una freqüència mitjana d'un marcador per cada dues frases. Per altra banda, Marcu (1998) indica que un algorisme d'anàlisi basat en la RST no pot explicitar les relacions suficients o amb una profunditat suficient per obtenir mesures determinants de la rellevància dels elements del text. Sembla evident que caldria recórrer a altres paràmetres per la millora de les mesures de rellevància obtingudes mitjançant l'estructura retòrica, sense renunciar però a la independència de domini que ens proporciona.

2.3 La postura que prenem

El nostre objectiu és plantejar un sistema de resum automàtic per a l'espanyol que ha de superar algunes de les mancances principals dels sistemes de resum actuals: principalment, la manca d'adequació a l'objecte que tracten els sistemes basats en mètodes estadístics, per altra banda, la dependència de domini i l'alt cost computacional que presenten les aproximacions que incorporen un coneixement lingüístic important. És per això que proposem d'adaptar els treballs d'Ono, Miike i Sumita (1994) i de Marcu (1997) a l'espanyol. Sembla que l'aproximació d'aquests investigadors supera de manera força satisfactòria les mancances que acabem d'exposar, perquè

- presenta **independència de gènere i domini conceptual**, ja que explota característiques que defineixen els textos i, per tant, són presents en tots ells
- aporta **coneixement lingüístic** per donar compte dels textos sense restricció de domini, cosa que, en l'estat de desenvolupament actual de les tècniques de PLN, implica una aproximació superficial al processament del llenguatge.
- com a conseqüència d'una aproximació superficial al processament del llenguatge, presenta un **cost computacional baix**, de manera que s'adapta a les nostres limitacions pel que fa a recursos de PLN
- respecta la **dimensió documental** dels textos, ja que incorpora allò que les ciències documentals diuen que cal saber sobre els textos com a documents: els textos són estructures d'informació, i sovint responen a patrons regulars que cal detectar i sistematitzar. Reconèixer aquestes estructures equival a reconèixer l'organització de les idees en un text, que és l'objectiu final de l'anàlisi retòrica dels textos.
- explota característiques lingüístiques de **nivell textual i discursiu** i mecanismes lingüístics d'organització de la informació, i adapta per a ús computacional línies de recerca, com la pragmàtica i l'anàlisi del discurs, poc formalitzades
- resulta **cognitivament plausible**, ja que es basa en el supòsit que existeixen mecanismes cognitius de processament del llenguatge que es reflecteixen en els textos (Hobbs 1985), i que podem detectar i sistematitzar aquestes marques per accedir a la informació associada a aquests mecanismes i explotar-la per realitzar una certa comprensió dels textos

D'aquesta manera, pretenem arribar a un resum satisfactori tant pel que fa a la seva adequació al text original com per la seva bona formació com a text autònom: que reculli les idees principals del text original, és a dir, que pugui determinar quins són els seus continguts, que reflecteixi l'organització de les idees en el text, de manera que pugui determinar quins constitueixen l'espina dorsal que articula el discurs, és a dir, que pugui determinar quines són les idees més importants d'un text, i que exposi tot això de manera coherent.

2.3.1 Retòrica i Argumentació

Com ha quedat de manifest, el nostre objectiu principal és donar compte dels continguts d'un text i la seva organització. En aquest sentit, Dressler i Beaugrande (1997: 50) apunten que el principal aspecte comú dels estudis clàssics en retòrica i la jove disciplina de l'anàlisi del discurs és que ambdós suposen que hom pot "`controlar de manera sistemàtica l'accés a les idees i a la manera d'organitzar-les`", cosa que s'adiu perfectament amb el nostre objectiu prioritari. És més: en l'evolució dels treballs sobre les relacions de coherència en els textos, sobretot a partir de la proposta de la RST, la nomenclatura de 'relacions retòriques' s'ha generalitzat en la bibliografia sobre el tema. També en la Teoria de l'Argumentació (Ducrot i Ascombre 1994), en què es basen la major part d'estudis sobre marcadors discursius de l'espanyol, es dóna compte d'un tipus de relacions textuals similars. És per tot això que considerem adequat anomenar 'relacions retòriques' els fenòmens textuals que ens permeten donar compte de com reconeixem els conceptes recollits en un text i la seva organització.

Com Hobbs (1985), partim dels supòsits que els textos poden ser no atòmics, que els textos no atòmics es divideixen en segments, els quals estableixen relacions entre ells, suposem que podem donar compte d'aquestes relacions en una estructura i que aquesta estructura és part constituent del text com a objecte lingüístic i comunicatiu unificat. Proposem de detectar les marques superficials de l'estructura retòrica per tal de reconèixer de manera automàtica la manera com s'organitzen les diferents parts d'un text. Aquesta aproximació ens aporta una gran independència de gènere i domini i un cost computacional baix, a canvi però d'un tractament certament superficial dels textos. Ara bé: malgrat que el processament dels textos que volem dur a terme es basa en marques de la seva estructura superficial, el coneixement lingüístic associat a aquestes marques no ho ha de ser gens, de superficial.

Ens proposem d'adaptar la proposta d'Ono, Miike i Sumita (1994) i, especialment, de Marcu (1997)[8], al processament de l'estructura retòrica dels textos en espanyol de cara a obtenir-ne una representació adequada i suficient per elaborar-ne el resum de manera automàtica. Com ja hem apuntat, també el treball de Jing (1998) fa aportacions significatives per a explotar les marques d'estructura textual, així que el tindrem en compte a l'hora de plantejar la nostra aproximació. Ara bé, no podem oblidar que aquests investigadors han treballat sobre el japonès i l'anglès respectivament, i que, malgrat existeixin característiques sobre l'estructura retòrica i l'argumentació comunes a la major part de llengües, també és cert que cada llengua realitza de manera idiosincràtica els mecanismes de processament del llenguatge i de coherència discursiva generals. És per això que tenim molt i molt en compte els estudis sobre les relacions discursives en espanyol, parant especial atenció als que se centren en donar compte del funcionament de les marques explícites d'aquestes relacions (vegeu secció 3). Donat que la major part dels treballs sobre marcadors del discurs en l'àmbit espanyol es basen en la teoria de l'Argmentació (Ascombre i Ducrot 1994), i donat que, en línies generals, aquesta teoria no presenta discrepàncies de consideració amb la RST, sinó que dóna compte de fenòmens similars des d'una perspectiva lleugerament diferent, hem intentat integrar els elements d'ambdues propostes que més adequats ens han semblat per les nostres limitacions, possibilitats i objectius.

D'aquesta manera, hem pres els principis bàsics d'organització dels textos sobre els quals es basteix la RST i els hem integrat amb informació de marcadors del discurs en espanyol, organtizada segons els paràmetres de la Teoria de l'Argumentació. Tenim en compte les següents característiques sobre l'estructura textual:

nuclearitat la relació bàsica que dóna lloc a la jerarquització de les relacions retòriques de la RST. Sembla que aquesta relació recull un fenomen molt general del llenguatge: el fet que els objectes lingüístics estan formats per constituents i que aquests constituents es poden relacionar entre ells de manera jeràrquica (Bloomfield 1933). Ara bé, malgrat que puguem

[8] Ens basem especialment en Marcu perquè la seva proposta ens és més transparent per qüestió de la llengua.

considerar la hipotaxi com un universal lingüístic, les normes que regulen l'organització dels constituents són particulars per cada nivell lingüístic. Sovint s'ha considerat la hipotaxi discursiva com una extensió directa de la subordinació oracional, sense tenir en compte les característiques específiques del text. És per això que el concepte de nuclearitat, centrat en la dimensió discursiva del llenguatge, resulta especialment adequat per donar compte de les estructures textuals.

orientació argumentativa: aquest paràmetre, procedent de la Teoria de l'Argumentació, contribueix a evidenciar la línia argumentativa d'un text, ja que dóna compte de les relacions entre segments segons si prenen la mateixa direcció argumentativa, és a dir, si un dels segments reprèn i continua una idea o una postura donada en un altre segment, o si, per contra, un segment aporta informació que s'oposa a allò dit en un altre. Aquest tipus d'informació pot ser molt valuosa a l'hora de determinar quins són els segments que basteixen l'espina dorsal d'un text.

força argumentativa: aquest paràmetre, també desenvolupat per la Teoria de l'Argumentació, contribueix a donar compte de la distribució de la rellevància en els textos, ja que, donats dos segments de text amb orientacions argumentatives diferents, determina quin dels dos és el que s'imposa i, per tant, quin és més central en el desenvolupament del text i quin, per contra, és rellegat a la categoria d'argument marginal. Creiem que aquesta informació contribueix a determinar amb més seguretat les relacions de nuclearitat.

La RST dóna molta importància al contingut retòric de les relacions entre segments de text. No obstant, donat el cost que requereix l'adquisició d'aquest coneixement sobre els textos, hem considerat més viable d'explotar la informació sobre l'estructura retòrica dels textos, deixant de banda el contingut. Així, en un primer moment de la nostra investigació, tractarem de donar compte del vessant purament estructural de les relacions entre segments de text. Creiem que mitjançant els tres conceptes que acabem d'enunciar podrem determinar la distribució de la rellevància en els textos en un nivell superficial. Com mostra en part el treball de Jing (1998), aquesta informació resulta molt útil per a l'elaboració d'un resum del text, i el seu cost de desenvolupament ens és assequible. No obstant, esperem que en un estadi més avançat del nostre treball podrem incorporar també el contingut retòric de les relacions, ja que aquesta informació pot enriquir considerablement l'anàlisi que obtinguem del text.

Seguint la proposta de la RST i el treball posterior que Marcu ha dut a terme, sembla clar que, si podem donar compte de cert contingut de les relacions entre segments, podrem arribar a una representació més acurada dels textos, que ens aportarà informació útil per elaborar el resum d'un text. Ara bé, tractar de manera adequada el contingut retòric de les relacions textuals requereix un treball considerable. En primer lloc, caldrà que considerem diversos conjunts de relacions retòriques proposats en la bibliografia sobre el tema, amb l'objectiu de delimitar el tipus de continguts que tractarem, tenint en compte la seva utilitat per al nostre objectiu i també el fet que puguem tractar-les donades les nostres limitacions de PLN. Un cop delimitades aquestes relacions, serà convenient elaborar una tipologia operativa que ens permeti de donar compte de la informació discursiva de què podem disposar. A través d'aquesta tipologia també podrem organitzar els marcadors del discurs, establint diversos nivells d'especificitat, grups de comportament, mecanismes d'herència entre nivells de la tipologia, etc.

Així, creiem que si establim una tipologia de relacions retòriques i associem a cada marcador un contingut retòric concret, podrem arribar a determinar de manera automàtica les relacions que s'estableixen entre els segments de text que connecta aquest marcador. Aquestes relacions contribuiran, entre altres coses, a determinar la línia argumentativa del text, el fluxe d'informació, les relacions de dependència entre els segments de text i la distribució de rellevància en aquests segments.

2.3.2 Proposta per una aproximació al resum automàtic

A partir d'aquest estudi previ, hem dissenyat una proposta per abordar el resum automàtic de textos sense restricció de domini en espanyol. La nostra proposta es basa en l'anàlisi de l'estructura retòrica, per tal d'obtenir una representació del text a partir de la qual es pugui obtenir automàticament un resum. Probablement, aquesta representació del text no serà suficient per ella mateixa per constituir un sistema de resum automàtic independent. Seguint la tendència majoritària en els sistemes de RA, és molt possible que l'anàlisi retòrica dels textos que s'integri en un sistema modular de resum en què col·laborin diversos tipus d'informació sobre els textos.

Per desenvolupar la nostra proposta, hem trobat que cal:

1. **delimitar l'objecte d'estudi**, és a dir, definir les unitats de treball l'anàlisi de l'estructura discursiva dels textos: els segments de text, les relacions entre els segments de text i les marques d'estructura discursiva

   (a) **relacions retòriques**: partim de la proposta de Marcu (1997), però considerem que cal adequar-la a l'espanyol[9] i, més concretament, als nostres objectius i limitacions. En un primer moment, treballarem tan sols el vessant més estructural de les relacions retòriques, sense explotar-ne el contingut, en un punt intermig entre el treball de Marcu (1997) i Jing (1998).

   (b) **marcadors del discurs**: hem de delimitar el tipus de marques d'informació discursiva que explotarem. Com que no tractarem el contingut retòric fins més endavant, en un primer moment ens centrarem en aquelles marques que evidencien relacions purament estructurals en el text, és a dir, relacions de nuclearitat entre segments. Per fer-ho, comptem amb la proposta de Marcu, de caire computacional, i amb els diversos treballs que sobre partícules discursives en l'àmbit espanyol, que aporten una perspectiva teòrica gens menyspreable. No obstant, manca una proposta de tractament computacional dels marcadors del discurs per a l'espanyol, així com un recull exhaustiu d'aquest tipus de marques.

   (c) **segments discursius**: trobem un buit teòric important pel que fa a la definició del concepte de segment com a unitat mínima de discurs en un nivell de granularitat alt. Donat que es tracta d'una unitat de treball fonamental, trobem necessari definir-la adequadament. Una definició clara d'aquest concepte ha de facilitar la comunicació científica i, a més, una noció sòlida de segment del discurs es perfila com una contribució important per a la definició d'altres elements discursius, com els marcadors i les relacions.

2. **detectar en els textos** aquella informació útil per al nostre objectiu que estem en disposició de tractar donades les nostres limitacions de PLN i sistematitzar-la per al seu ús en el processament computacional del discurs per a resum automàtic. Això equival a detectar i sistematitzar les marques d'estructura retòrica, tant aquelles que evidencien relacions com les que assenyalen les unitats entre les quals s'estableixen aquestes relacions.

3. **desenvolupar les eines i recursos** que ens permetin d'explotar aquestes marques d'estructura dels textos per tal d'obtenir una representació del text que aporti informació útil per elaborar-ne un resum

4. donar compte de com l'anàlisi de **l'estructura retòrica pot contribuir al resum** automàtic d'un text i desenvolupar un conjunt d'heurístiques per elaborar un resum a partir d'aquesta informació

[9] Les relacions retòriques de la RST són flexibles precisament per facilitar la seva adaptació a tasques i característiques lingüístiques concretes.

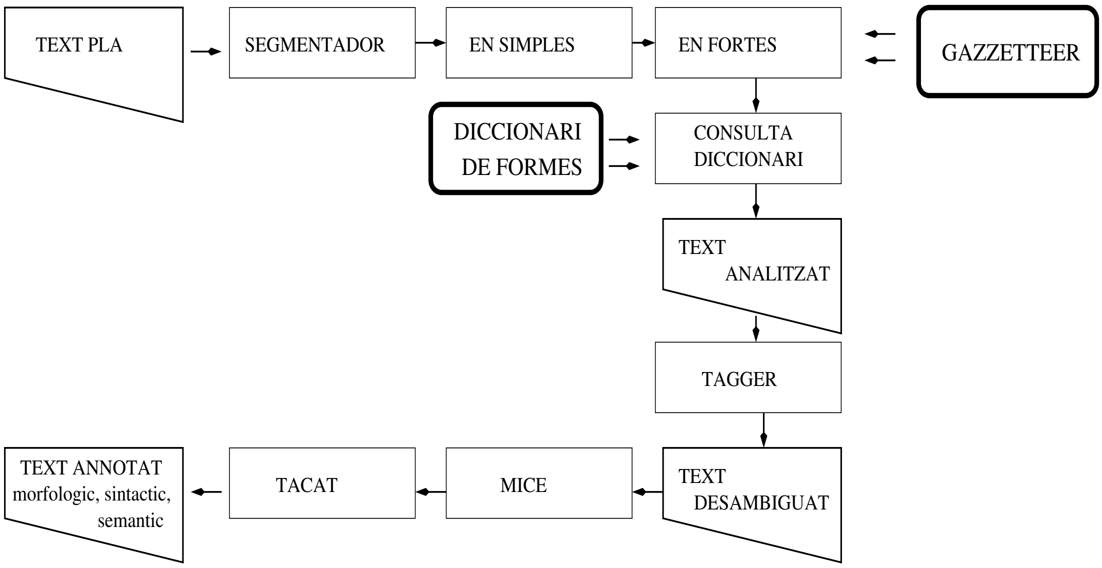

Figura 2.5: Arquitectura del procés d'anàlisi morfosintàctica TALP-CLiC

(a) heurístiques que explicitin com l'estructura retòrica d'un text determina quins són els segments més **rellevants** per a un resum

(b) heurístiques d'**integració** d'informacions heterogènies sobre els textos: quantitatives, documentals (posició, paraules clau de contingut d'un domini, etc.), retòriques, narratives, etc.

(c) heurístiques d'**elaboració de resums** a partir d'una representació estructurada d'un text que integri diversos tipus d'informació textual: extracció de fragments rellevants, resolució de problemes de coherència, compressió del text i generació del resum

Pel que fa al desenvolupament efectiu de la nostra proposta, creiem que els resultats poden millorar si treiem profit dels recursos de PLN de què disposem en cadascun dels nivells lingüístics, per tal d'obtenir el màxim d'informació pertinent dels textos:

1. anàlisi **morfosintàctic**: utilitzarem tres eines que treballen conjuntament per obtenir l'anàlisi morfosintàctic de text lliure en espanyol (Atserias et al. 1998b) (vegeu figura 2.5):

   (a) MACO (Arévalo et al. 2001), analitzador morfològic de l'espanyol

   (b) RELAX (Padró 1997), desambiguador d'etiquetes morfològiques a partir d'aprenentatge automàtic

   (c) TACAT (Atserias et al. 1998a), analitzador sintàctic superficial

2. detecció **d'entitats**: es troben en estat de desenvolupament en el grup de recerca TALP-CLiC eines i recursos per a la detecció de les entitats amb nom dels textos (Arévalo i Simón 2000) (vegeu figura 2.1):

   (a) reconeixement d'entitats simples i fortes mitjançant una base de dades d'entitats amb nom (GAZZETTEER)

   (b) MICE, mòdul d'identificació i classificació d'entitats

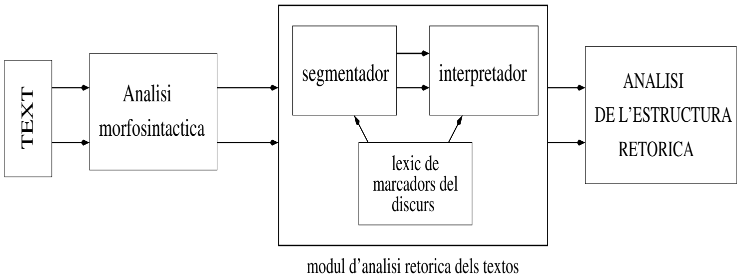

Figura 2.6: Disseny d'un mòdul d'anàlisi retòrica dels textos

En un text anotat amb informació morfosintàctica i d'entitats, aplicarem un procés d'anàlisi retòrica, en dos passos (vegeu figura 2.6):

1. **segmentació retòrica** mitjançant un lexicó de marcadors del discurs (en desenvolupament)
2. **interpretació** retòrica mitjançant un lexicó de marcadors del discurs (desenvolupament futur)

A partir de l'anàlisi retòrica del text, caldrà aplicar heurístiques que explotin aquesta informació per a resum automàtic. Probablement, aquestes heurístiques hauran de contemplar també la interacció entre aquesta i altres tipus d'informació útils per a resum automàtic (paraules clau de contingut, qüestions de format documental, etc.). També caldrà dur a terme una avaluació de cadascun d'aquests passos del sistema, per la qual cosa s'hauran de dissenyar i dur a terme mètodes d'avaluació de segmentació retòrica, d'interpretació retòrica i de resum automàtic.

Per tant, la nostra tasca principal consistirà en el disseny i desenvolupament de les eines i recursos de PLN per l'anàlisi retòrica i, en un segon moment, per al resum automàtic basat en l'estructura retòrica dels textos. Tot seguit exposem l'estudi i desenvolupament d'una primera versió de la base de dades de marcadors del discurs, que conté una mostra de marques que tenen associada la informació discursiva necessària per al seu funcionament en el sistema. En segon lloc, presentem el segmentador retòric, actualment en procés de millora de resultats. Aquesta eina utilitza part de la informació que hem introduït en la base de dades de marcadors del discurs.

Capítol 3

Els Marcadors del Discurs

Hem exposat ja la vàlua de les marques superficials de l'estructura discursiva per al processament del discurs, parant especial atenció a la seva utilitat per elaborar un resum de manera automàtica. Tot seguit presentem algunes qüestions d'interès per al processament dels marcadors del discurs en lingüística computacional, des de les dues perspectives que ens interessen: la teòrica i l'aplicada. Com veurem, la informació que proporciona i requereix l'estudi dels marcadors des d'aquestes dues perspectives és complementària, per tant, cal integrar-les totes dues per tractar de manera satisfactòria i adequada els marcadors del discurs.

Pel que fa a la terminologia, és gairebé tan diversa com diverses són les aproximacions a l'estudi d'aquestes unitats: es parla des de 'cue phrases' (Passonneau i Litman 1993, Knott 1996, entre altres), 'clue phrases' (Cohen 1987), 'discourse cues' (Knott 1996), 'connectors clausals' (Knott i Mellish 1996), 'connectors discursius' o simplement 'connectors' (Webber et al. 1999, Montolío 2001, entre altres), 'enllaços extraoracionals' (Gili Gaya 1943, Fuentes 1987), 'operadors discursius' (Casado Velarde 1991), 'partícules discursives' (Montolío 2001) i també 'marcadors discursius' (Schiffrin 1987, Marcu 1997, entre altres). Nosaltres hem adoptat aquest darrer terme seguint el plantejament de Marcu, però durant aquest treball utilitzarem els termes 'marca de discurs' i 'marca discursiva' per referir-nos a informació d'estructura textual en un nivell superficial, i emprarem 'marcador' o 'marcador del discurs' per referir-nos al tipus d'element lingüístic que ara ens ocuparem de delimitar.

Hem trobat que la vessant lingüística dels marcadors del discurs espanyols està a bastament desenvolupada, però no podem dir el mateix de la computacional. És per això que hem hagut de fer una exploració i sistematització dels aspectes més formals d'aquestes unitats lingüístiques, una tasca que hem subordinat a les necessitats sorgides de la proposta que hem fet en el capítol anterior. En la segona part d'aquest capítol presentem un estudi de corpus sobre marques discursives per al seu ús en el processament dels textos, una proposta purament operativa.

3.1 El processament dels marcadors del discurs

Els estudis sobre marcadors del discurs han crescut en gran mesura a partir del gir paradigmàtic en lingüística que va suposar la perspectiva pragmàtica, la lingüística del text i l'anàlisi del discurs, que va posar en evidència el paper clau d'aquestes peces lingüístiques en l'organització del discurs, i des d'aleshores s'han dedicat importants esforços a determinar la seva natura i descriure el seu comportament. No obstant, encara som molt lluny del consens teòric sobre aquests elements i la seva funció, com fa palès la síntesi bibliogràfica de Pons (1998: 41-47). Això dificulta enormement la seva formalització i, per tant, la seva aplicació a usos computacionals.

Com hem dit abans (2.2), en el discurs trobem informació molt diversa, per tant, com podem trobar marques explícites d'informació discursiva de molts tipus:

- mecanismes necessaris per comprendre el discurs a partir del contingut proposicional explícit: **inferències** lògiques i **implicatures** discursives (Grice 1975)
- **estats mentals** dels participants en l'acte comunicatiu: creences, opinions, predisposicions (Wing i Scholnick 1981)
- **estats atencionals** dels participants, cosa que delimita el marc de referència en què es troba el discurs (Grosz i Sidner 1986)
- **evolució temàtica** del discurs, canvi o continuació de tema
- **estructura argumentativa** del discurs (Anscombre i Ducrot 1983)
- i molts més, segons quina sigui la funció pragmàtica que s'estudiï, el nivell d'anàlisi del text, etc.

Però, com en el cas de les aproximacions al tractament del discurs, són pocs els que tracten totes les facetes possibles dels marcadors alhora. Tot seguit exposarem alguns treballs que aprofundeixen alguns dels aspectes que considerem més útils per a l'ús de les marques explícites d'estructura discursiva en l'anàlisi de l'estructura retòrica dels textos.

3.1.1 Delimitació del concepte

És clar que l'existència en les llengües de partícules amb un significat i funció bàsicament discursiu és conegut des d'antic. Ara bé, potser precisament pel seu caràcter d'elements marginals en l'oració (Martín Zorraquino i Portolés 1999), tradicionalment se n'ha fet un estudi també marginal. És per això que, durant molts anys, els marcadors havien estat llistats de manera poc sistemàtica, o senzillament mencionats de manera anecdòtica, en les gramàtiques i els tractats d'estilística. També és cert, però, que alguns autors han dedicat una atenció considerable a aquest tipus d'unitats lingüístiques.

En llengua espanyola, comptem amb una llarga tradició d'estudi sobre la natura retòrica dels textos (una bona mostra és Salinas 1541). En canvi, són menys els gramàtics, i encara menys els lexicògrafs, que han estudiat i intentat sistematitzar les marques d'estructura del discurs de l'espanyol. Garcés (1791) és una bona mostra de l'abordatge tradicional d'aquest tipus d'unitats lingüístiques: en una monografia sobre les partícules de la llengua castellana, llista per ordre alfabètic un bon nombre de paraules categorialment heterogènies, en dóna exemples d'ús amb textos d'autors clàssics i comenta les funcions pragmàtiques associades a cadascuna (valors expressius, intencions del parlant, etc.). Per tant, el tractat de Garcés constitueix una font d'informació inestimable per a tots els lingüistes posteriors, tant per la gran quantitat d'exemples que aporta com per la varietat d'elements textuals que tracta: Bello, per exemple, inclou en la seva *Gramàtica* (1847) una llista molt similar a la de Garcés (per una comparació de totes dues llistes vegeu Lliteras 1995).

És ja en el marc de la lingüística moderna que la gramàtica espanyola comença a tractar de manera més sistemàtica els marcadors discursius: Gili Gaya (1943), parlant dels 'enllaços extraoracionals', dóna tota una sèrie de característiques definitòries de les partícules amb una funció discursiva: el fet que estan vinculades amb elements extraoracionals, que presenten una gran variabilitat categorial i distribucional i, alhora, són invariables, i, sobretot, el seu significat de 'coherència discursiva'. Alcina i Blecua (1975) parlen de 'ordenadors del discurs' i 'ordenadors lèxics', tipus de partícules que quedarien inclosos en els 'enllaços extraoracionals' de Gili i Gaya. Amb l'arribada de la perspectiva pragmàtica trobem un salt quantitatiu i qualitatiu en els estudis sobre marcadors del discurs, com bé ho mostra el recull de Montolío i Martín (1988) i el compendi bibliogràfic de

Cortés (1995a i 1995b). Trobem dues línies principals d'investigació sobre marcadors en espanyol: la que aplica teories d'anàlisi sobre text escrit (Montolío 1997, Portolés 1998a) i la que estudia el vessant més dialògic d'aquestes marques (Briz 1993a i 1993b, Pons Bordería 1998), a banda de múltiples estudis sobre marcadors particulars i perspectives concretes.

Cal destacar el treball de Martín Zorraquino i Portolés (1999), que descriuen gairebé dos-cents marcadors del discurs des de la perspectiva de la Teoria de l'Argumentació, però sense deixar de banda les aportacions d'altres punts de vista, com, per exemple, una perspectiva històrica del tractament d'aquestes unitats lingüístiques en la lingüística hispànica. La classificació de marcadors que proposen és una de les més acceptades en la bibliografia sobre el tema: distingeixen entre

estructuradors de la informació: comentadors, ordenadors, digressors

connectors: additius, consecutius, contraargumentatius

reformuladors: explicatius, rectificatius, de distanciament, recapitulatius

operadors argumentatius: de reforç argumentatiu

conversacionals: de modalitat epistèmica, de modalitat deòntica, enfocadors d'alteritat, metadiscursius

Ara bé, el que sens dubte constitueix la seva principal aportació és la formulació d'una definició de marcador del discurs, a la qual ens sentim molt pròxims (Martín Zorraquino i Portolés 1999 : 4057):

```
Los 'marcadores del discurso' son unidades lingüísticas invariables, no ejercen una
función sintácica en el marco de la predicación oracional -son, pues, elementos marginales-
y poseen un cometido coincidente en el discurso: el de guiar, de acuerdo con sus
distintas propiedades morfosintácticas, semánticas y pragmáticas, las inferencias
que se realizan en la comunicación.
```

També presenten una interessant discussió de les propietats que definirien un marcador del discurs seguint aquesta definició:

- no tenen un significat conceptual, sinó de **processament**, és a dir, no en podem predicar cap valor de veritat ni contribueixen de cap manera al valor de veritat de les oracions circumdants
- presenten una gran **movilitat** pel que fa a posicions sintàctiques, poden situar-se en membres del discurs que constitueixen categories lèxiques i sintagmàtiques diverses, però no obstant podem donar compte de tendències pel que fa a marques d'**entonació** o de **puntuació**
- tenen una **relació sintagmàtica** amb la totalitat del sintagma en què es troben, fins i tot quan es tracta d'una oració
- són **invariables**
- no poden tenir modificadors ni complements, no es poden coordinar amb un altre marcador ni poden ser negats, i tampoc no poden ser destacats amb perífrasis de relatiu

Com hem pogut comprovar, les múltiples aproximacions actuals als marcadors del discurs en espanyol contribueixen a donar-nos una perspectiva prou rica i àmplia d'aquests fenòmens lingüístics. Però malgrat aquest panorama engrescador, trobem dues mancances principals perquè puguem fer ús d'aquesta informació en un sistema de PLN:

1. no existeix encara una obra de referència que doni compte d'un conjunt **exhaustiu** de les partícules discursives en espanyol. Donat que aquesta mancança és fa cada cop més evident en diverses àrees de l'aplicació de la llengua (l'ensenyament de l'espanyol com a llengua estrangera, la descripció gramatical,...) esperem que aviat podrem comptar amb una obra que cobreixi aquesta mancança[1]

2. fins al moment, no coneixem cap estudi que tracti els marcadors del discurs en llengua espanyola de manera **formal** i sistemàtica, i encara menys com a font d'evidència per a l'extracció automàtica d'informació de l'estructura textual. Aquest coneixement és bàsic per al desenvolupament d'una aplicació d'anàlisi retòrica dels textos que es basi en marcadors del discurs. Per tant, caldrà detectar, descriure i sistematitzar les marques lingüístiques superficials que ens aporten informació rellevant per al nostre objectiu. Aquestes marques no necessàriament han de coincidir amb els marcadors del discurs tal com s'entenen en lingüística teòrica

En canvi, alguns estudis sobre partícules discursives en llengües diferents de l'espanyol presenten un grau de formalització considerable. El punt de referència bàsic en l'estudi formal dels marcadors és Schiffrin (1987), malgrat que el seu enfocament socio-lingüístic podria fer-nos pensar que es tracta d'una aproximació poc formal. Schiffrin proposa un model de coherència discursiva en què distingeix cinc nivells: estructura d'intercanvis, estructura d'acció, estructura ideacional, marc participatiu i estat informatiu. Partint d'aquest model, descriu un petit conjunt de marcadors en el discurs parlat, determinant per cadascun d'ells una funció primària pel que fa als torns de parla (estructura d'intercanvis), als actes de parla (estructura d'acció), a l'estructura proposicional (estructura ideacional), a les relacions entre els participants (marc participatiu) o a la informació (estat informatiu). No obstant, Schiffrin també contempla la possibilitat que un marcador estigui associat a més d'un d'aquests nivells, és més, considera que aquest funcionament en múltiples nivells simultanis dóna coherència al discurs, perquè interrelaciona diversos tipus d'informació que es complementen.

El treball de Schiffrin presenta una perspectiva molt reveladora sobre els marcadors del discurs, i constitueix un punt de referència per als sistemes de PLN que exploten la informació discursiva. No obstant, la principal vàlua d'aquesta proposta és alhora el principal obstacle per la seva implementació en un sistema computacional: el fet que tracta tipus molt heterogenis d'informació discursiva, interrelacionant-los, dóna una complexitat i una profunditat d'anàlisi que presenten importants dificultats d'aplicació amb les tècniques de PLN actuals. No obstant, Schiffrin aporta una perspectiva molt enriquidora sobre aspectes del discurs que, segons el nostre parer, han de tendir a la integració en tant que els avenços en LC, PLN i IA ho permetin.

Un altre treball a tenir molt en compte és el de Knott (1995). Knott fa una important revisió dels treballs que es basen en les relacions de coherència dels textos, i troba un punt feble en el fet que apliquen mètodes que podríem considerar deductius per donar compte de com es realitzen aquestes relacions en els textos. En la major part dels casos, es sistematitzen les marques d'informació discursiva a partir d'un conjunt de relacions determinades a priori, sovint preses d'alguna teoria de la coherència textual (sobretot la RST). Knott creu que, amb una metodologia tan deductiva, es pot arribar a representacions de les marques discursives fortament esbiaixades pels presupòsits inicials, si bé és cert que en el treball de sistematització sovint es fan adequacions a la realitat textual. No obstant, també aquest darrer mecanisme d'adaptació a la realitat textual pot prestar-se a la manca de rigor científic, ja que es poden produir explicacions *ad hoc,* que solucionin aspectes puntuals però sense mantenir una coherència amb el cos de la teoria.

[1]Actualment s'està duent a terme un projecte de DGICYT anomenat 'Diccionario de Partículas Discursivas del Español', de què formen part A. Briz, E. Montolío i J. Portolés.

Knott proposa invertir aquest mètode per eliminar tot esbiaixament possible, és a dir, partir de les marques discursives per elaborar una tipologia de relacions retòriques neutra, dirigida per l'evidència textual i no per cap teoria.

La metodologia que proposa Knott es basa en dues passes principals:

- **identificació** de marcadors: proposa un test per determinar quines unitats lingüístiques són marcadors, en tres passes:

  1. detecció d'un candidat a marcador
  2. aïllament de la frase en què es troba sintàcticament inclòs el candidat a marcador, substituïnt tot element anafòric o catafòric pels seus antecedents i incloent tots els elements elidits[2]: si la frase resulta discursivament completa i feliç, no es tracta d'un marcador, si la frase resulta incompleta, es passa al següent pas.
  3. eliminació del candidat de la frase en què es troba: si la frase resulta discursivament completa i feliç, es tracta d'un marcador.

  exemple:

  1. detecció del candidat:
  ... *John and Bill were squabbling: John was angry* because *Bill owed him money. That was how it all started...*
  2. aïllament del candidat:
  because *Bill owed John money.*
  La frase resulta incompleta, per tant, passem al següent pas.
  3. eliminació del candidat:
  Bill owed John money.
  La frase resulta completa, per tant, es tracta d'un marcador.

  Knott (1995: 71)

- **classificació** de marcadors en una tipologia: mitjançant el mecanisme bàsic de la commutació, determina la distribució dels marcadors en els textos, de manera que pot establir relacions de sinonímia, hiponímia, hiperonímia, exclusió i substitució contingent entre marcadors. Aquestes relacions s'estableixen de manera ideal per cada parell de marcadors, tot i que a la Knott tan sols ha donat compte dels parells més informatius.

Amb aquest treball, Knott cobreix una mancança important en l'estudi dels marcadors del discurs: en primer lloc, proposa mecanismes científicament rigorosos per la identificació i sistematització dels marcadors del discurs, per tant, aporta una delimitació força formal de l'objecte d'estudi. En segon lloc, constitueix un recurs lingüístic de primer ordre tant pel que fa a la quantitat de marcadors que recull (150) com per l'acurada i exhaustiva descripció que en fa. I, en tercer lloc, basteix una tipologia de marcadors de l'anglès que sembla demostrar la validesa d'aquests mecanismes. És per això que aquest treball constitueix un sòlid punt de referència en l'estudi formal dels marcadors del discurs i ha estat utilitzat en diverses aplicacions computacionals posteriors (Marcu 1997, Kruij -Korbayova i Webber 2000).

En conclusió, hem vist com les diverses aproximacions a l'estudi dels marcadors contribueixen a delimitar aquest concepte de múltiples maneres. Per una banda, determinen de quin tipus de lèxic es tracta. Així, descriuen la seva natura a mig camí entre paraules lèxiques, amb contingut i susceptibles de ser taxonomitzades, i paraules gramaticals, amb una funció preeminent de processament del discurs. Aquesta natura indefinida queda reflectida també en qüestions polèmiques com la invariablitat dels marcadors del discurs, el tipus de significat que poden expressar o la manera més adequada de donar-ne compte.

[2]Knott considera que es donen marcadors complexos que poden contenir un element anafòric en si mateixos, en aquest cas, no es restitueix l'element anafòric, sinó que es considera part del marcador i per tant es manté invariable durant tot el test.

Així doncs, trobem que interessos molt diversos dirigeixen aproximacions també molt diverses sobre tipus heterogenis d'unitats englobats sota el concepte genèric de marcador del discurs. Cada aproximació acota el tipus d'unitat a tractar pel seu significat i comportament, però no trobem una delimitació que puguem considerar general del concepte. És per això que també nosaltres haurem de proposar la nostra pròpia delimitació de l'objecte, d'acord amb els nostres objectius i necessitats. No obstant, els diversos estudis sobre marcadors ens aporten informació que ens pot ser molt útil, i que hem classificat en:

- conjunt d'**unitats susceptibles de ser marcadors** llistades per forma: encara que hi ha diversos reculls de marcadors, no coneixem cap aproximació que tingui caràcter exhaustiu, sinó que la manera més habitual de descriure el conjunt d'aquests elements és a través de diversos criteris per identificar marcadors, és a dir, no per extensió sinó per intensió. Un punt de divergència és la **invariabilitat** dels marcadors: mentre Martín Zorraquino i Portolés declaren categòricament que es tracta d'elements invariables, Knott dóna regles de composició de marcadors que donen compte de mecanismes productius en la creació de nous marcadors
- **funcions** discursives que realitzen:
  - modificacions del contingut semàntic i proposicional dels objectes lingüístics
  - diversos significats que ajuden al processament del discurs per part dels participants:
    - licitació d'inferències
    - organització de l'estructura informativa: distribució dels temes en el discurs, senyals de canvi de tema, marques de focalització
    - direcció de la línia argumentativa
  - explicitació d'intencions comunicatives

3.1.2 Necessitats del tractament computacional

Com hem vist en l'apartat anterior, les diverses aproximacions als marcadors del discurs aporten informació interessant i útil sobre aquests elements, i, sobretot, han contribuït a acotar el concepte de marcador del discurs, tot i que amb aproximacions, i també termes, molt diversos. També cal destacar la creació d'un marc d'estudi per aquests elements fins ara marginals i mancats d'una metodologia que permetés treballar-los amb rigor científic. Ara bé, en l'àmbit que ens interessa, el de les partícules discursives en espanyol, manca un model formal implementable. Les descripcions dels marcadors espanyols que coneixem presenten els següents inconvenients per a la seva aplicació en un sistema de PLN:

- no són prou **completes**
  - no contemplen aspectes dels marcadors que lingüísticament poden semblar irrellevants però que són cabdals per al seu funcionament en un sistema computacional, com per exemple les instruccions de segmentació associades a un marcador, és a dir, quins són els límits del seu abast i quines són les marques formals que els assenyalen, si n'hi ha.

    exemple: *[Las semifinales acabarán el 12 de enero], [**aunque** el programa alargará la competición hasta el 19 en previsión de posibles aplazamientos], [**como** ha sucedido en las tres rondas celebradas].*

    En aquest exemple, podem distingir tres segments, dos d'ells caracteritzats per marcadors del discurs: 'aunque' i 'como'. Els signes de puntuació delimiten l'abast d'aquests marcadors i en deteninen la funció discursiva inambigua.
  - no són exhaustives, és a dir, no fan un recull del màxim nombre de partícules discursives, sinó que la tendència generalitzada és a donar un tractament aprofundit a alguns marcadors, per establir un model de treball aplicable a tota la resta, però sense tractar efectivament tota la resta

– no són prou **formals**, és a dir, no donen la informació associada als marcadors de manera clara i inambigua, ni la sistematitzen de manera que pugui ser traslladada a una base de dades per a ús computacional

Així doncs, si volem desenvolupar una aplicació de PLN que es basi en aquestes marques explícites d'informació discursiva, podem determinar dues **mancances** principals pel que fa a la informació que ens aporta la perspectiva lingüística sobre aquestes marques:

1. manca un enfocament computacional dels marcadors del discurs per l'espanyol:
   (a) **definir** què pot ser **marcador** des del punt de vista computacional, és a dir, què aporta informació que sigui computacionalment útil per identificar relacions textuals en un nivell superficial
   (b) **definir** els **paràmetres** necessaris per a la descripció formal d'aquests elements per al seu ús computacional, per això cal distingir els diferents nivells de funcionament dels marcadors (distribució de rellevància, contingut semàntic, etc)
2. manca un **recull exhaustiu** de les possibles partícules que poden funcionar com a marcadors des del punt de vista computacional i la seva sistematització en una base de dades que en possibiliti l'ús per aplicacions computacionals.
   (a) definir els factors d'**ambigüetat** dels marcadors tant pel que fa a la seva funció discursiva com al seu contingut retòric
   (b) explicitar les **relacions** entre marcadors, que pot ser útil de cara a realitzar generalitzacions sobre comportaments i a descriure els marcadors de manera més acurada

Així doncs, les aproximacions majoritàries, com a mínim en l'àmbit espanyol, no tenen gaire en compte la vessant computacional dels marcadors del discurs. Però quan hem buscat en l'àrea de PLN una definició més computacional del concepte de marcador, hem trobat que la major part de sistemes que treballen amb marcadors del discurs prenen una postura purament procedural, sense buscar una definició que delimiti el tipus d'objecte amb què treballen. Una tendència generalitzada consisteix a treballar directament a partir de les llistes de marcadors que han desenvolupat altres investigadors, adaptant-les als objectius o necessitats de cada aplicació particular, però sense donar-ne compte de manera sistemàtica (Marcu 1997, Kruij -Korbayova i Webber 2000). Així doncs, fins allà on arriba el nostre coneixement, la perspectiva computacional tampoc no ens ofereix una definició de marcador del discurs en què ens puguem basar per a desenvolupar el nostre sistema.

Respecte de l'exhaustivitat, considerem que és gairebé impossible de fer un recull de tots els elements lèxics de qualsevol tipus en una llengua, i, per tant, també és impossible de donar compte de tots els marcadors del discurs per l'espanyol. No obstant, com més ampli sigui el conjunt de marques amb què treballem, més informació tindrem per abordar l'anàlisi dels textos, per tant, trobem necessari anar més enllà dels reculls de les gramàtiques tradicionals i ampliar el nombre de marques discursives de què donem compte. Marcu (1997), per exemple, desenvolupa un sistema d'anàlisi retòrica i de resum automàtic amb un conjunt de 400 marcadors, tot i que assenyala la conveniència d'ampliar-lo per millorar l'efectivitat del sistema.

Per solucionar aquestes mancances de cara al desenvolupament del sistema proposat, hem dut a terme un estudi de marcadors que presentem tot seguit.

3.2 Estudi sobre corpus de Marcadors del Discurs

Com hem constatat en la secció anterior, manca per l'espanyol un recurs que reculli i sistematitzi les marques discursives per al seu ús computacional. Aquest recurs és un dels pilars del nostre objectiu final: el resum automàtic mitjançant l'anàlisi de l'estructura retòrica dels textos. És per això que ens hem proposat de desenvolupar aquest recurs. Ara bé: la base de dades de marcadors del discurs que hem elaborat no constitueix un recurs lingüístic de caire general, sinó que està marcadament orientada a l'aplicació que ens ocupa. Aquest objectiu ha determinat el tipus de marques que hi incloem i també la informació associada a cada element. Malgrat aquest plantejament principalment aplicat, creiem que la informació que hem recollit i sistematitzat pot ser ampliada i adaptada més endavant per a altres usos. D'aquesta manera, el recurs pot anar adquirint progressivament una dimensió més general.

3.2.1 Metodologia

Per tal de delimitar el nostre objecte d'estudi i desenvolupar el recurs projectat, hem dut a terme un estudi de marcadors del discurs sobre un corpus de la llengua espanyola. L'objectiu immediat d'aquest estudi és la detecció i sistematització de certes marques d'estructura en els textos, per al seu processament automàtic. Les nostres hipòtesis de partida es desprenen dels treballs de Ono et al. (1994) i Marcu (1997):

1. que en els textos es donen marques que expliciten certes relacions, que anomenem retòriques, entre els segments que constitueixen els textos
2. que, entre aquestes marques, n'hi ha de prou sistemàtiques per al seu processament amb mètodes automàtics superficials

En un primer estudi sobre corpus hem intentat validar la primera d'aquestes hipòtesis. L'adequació de la segona, però, tan sols es pot comprovar mitjançant el desenvolupament i aplicació d'eines de PLN que se serveixin efectivament d'aquestes marques i la informació que tenen associada. Així doncs, per validar aquesta hipòtesi hem desenvolupat una eina de segmentació basada en el reconeixement i utilització dels marcadors del discurs i de part de la informació retòrica que tenen associada.

En primer lloc, doncs, hem fet una prospecció d'un corpus molt ampli i representatiu de la llengua espanyola, per obtenir una impressió general del funcionament de les marques discursives. El nostre corpus de treball està format per:

LEXESP (5.000.000 paraules) el corpus del projecte LEXESP, una mostra representativa de diversos gèneres i registres de la llengua espanyola, obtingut gràcies al projecte LEXESP-III.

Vanguardia (6.500.000 paraules): una mostra de dos mesos del text complet del diari *La Vanguardia*, part del corpus CLiC-TALP.

Periódico (3.500.000 paraules): una mostra de dos mesos i mig del text complet del diari *El Periódico*, procedent del projecte NAMIC.

EFE (1.500.000 paraules): una mostra d'alguns dies de maig de notícies de l'agència de notícies EFE, procedent del projecte NAMIC.

Malgrat que aquesta primera aproximació ha estat molt intuïtiva, no podem ignorar la influència dels treballs sobre marcadors de l'apartat anterior. A partir d'aquí hem pogut comprovar que, efectivament, es donen marques superficials en els textos espanyols que n'expliciten l'estructura retòrica i argumentativa.

D'aquestes marques, moltes es corresponen amb els marcadors del discurs de què donen compte els lingüistes espanyols que han treballat en aquesta àrea, i són equivalents a la major part de les que utilitza Marcu (1997). Ara bé, hem observat que, tal com apunta Jing (1998), altres elements textuals, que no s'adiuen amb la definició de marcador proposada per Martín Zorraquino i Portolés (1999), contribueixen també a la detecció i delimitació dels segments de text i de les relacions intersegmentals que constitueixen l'estructura d'un text, per exemple, els signes de puntuació o algunes estructures sintàctiques. Com també assenyala Knott (1995), es donen alguns mecanismes molt productius de creació de marcadors, cosa que posa en qüestió la seva invariabilitat, com a mínim des del punt de vista computacional: per exemple, les construccions de sintagma preposicional amb pronom (p. ex.: *por eso, por esto, por ello, por lo que*), o moltes arrels amb un significat clarament retòric (p.ex..: *en concreto, más concretamente, concretando*). No obstant, per qüestions d'optimització de processament, hem donat compte dels marcadors com a ítems lèxics invariables, tot i que esperem poder desenvolupar un conjunt de regles que sistematitzin els mecanismes de producció de marcadors més generals.

Tampoc no hem donat compte d'altres fenòmens de cohesió textual que poden interactuar amb els marcadors, com la co-referència, la *consecutio tempore*, la negació, els elements circumstancials, etc. Aquests elements contribueixen també a fer del text una unitat lingüística cohesionada i dirigeixen en bona part el fluxe d'informació i la realització de les inferències necessàries per al processament del text. Ara bé, com ja hem dit abans, tot aquest tipus de fenòmens queden fora de l'abast de la nostra investigació actual.

Així doncs, ens proposem de tractar com a marques de l'estructura retòrica dels textos, essencialment, aquelles estructures sintàctiques, peces lèxiques i signes de puntuació que ens permeten delimitar els segments de text que funcionen com a unitats discursives mínimes o bé aporten informació sobre les relacions retòriques que estableixen aquests segments entre ells. Ens hem dedicat principalment a determinar i sistematitzar les característiques generals d'un grup de marcadors poc conflictius: pensem que, un cop aclarides les qüestions més generals respecte dels marcadors, i amb una descripció sistemàtica de les regularitats, podrem dedicar-nos a explicar les excepcions més notables. És per això que hem deixat de banda els marcadors que expliciten les postures dels participants en l'acte comunicatiu i, en general, aquells marcadors amb un comportament o un significat poc sistematitzable o poc clar. Hem constituït un primer conjunt d'uns 600 marcadors del discurs (vegeu Annex IV), recollits de diverses fonts:

- del formari de l'analitzador morfològic MACO (Arévalo et al. 2001) hem pres especialment aquells marcadors pertanyents a les tradicionals categories tancades de la llengua: conjuncions, algunes preposicions i locucions conjuntives i prepositives

- a partir de diverses fonts bibliogràfiques (Portolés i Zorraquino 1999, Marcu 1997, Knott 1995) hem tractat de delimitar el tipus d'informació discursiva que ens era més útil de tractar, de manera que hem recollit i sistematitzat els tipus de marcadors que tracta cada investigador i hem descartat els que ens resultaven poc útils per a resum automàtic o massa costosos en l'estat actual del nostre treball

- en una prospecció preliminar de corpus hem tractat d'identificar aquelles marques d'estructura discursiva que poden aportar-nos informació útil per a resum automàtic, és a dir, que poden contribuir a determinar la distribució de rellevància en els textos: bàsicament, algunes estructures sintàctiques i signes de puntuació

Per tal de determinar de manera clara el tipus d'unitats que tractarem, hem partit del concepte de 'significat de processament' que presenten els treballs sobre els marcadors del discurs. No obstant, nosaltres tractem un conjunt d'unitats força més ampli que el que proposen les aproximacions majoritàries, fortament determinats pel nostre objectiu computacional. Així, ens basem en la següent definició operativa de marcador del discurs:

marcador	ocurrències sense coma	ocurrències seguit de coma	ocurrències totals
y	202665	7452	210117
como	59843	144	59987
pero	30215	1054	31269
o	29365	706	30071
cuando	18979	557	19536
hasta	18574	19	18593
desde	17485	5	17490
porque	15979	832	16811
según	10976	3	10979
aunque	10355	220	10575

Taula 3.1: Marcadors més freqüents en corpus

> `tota unitat textual, lèxica, sintàctica o de puntuació, que explicita una relació retòrica concreta entre dos o més enunciats`

Un cop delimitat l'objecte d'estudi, hem dut a terme un segon estudi de corpus, per aprofundir en l'estudi de les marques seleccionades. Hem elaborat una mostra de treball, en què es troben representats tots els marcadors que hem estudiat. En aquesta mostra tractem la variació de domini però mantenim una certa homogeneïtat pel que fa al registre: formal i escrit, i pel que fa la gènere: periodístic. Així doncs, hem deixat de banda la variació del discurs literari, d'especialitat i col·loquial. Pensem que, d'aquesta manera, disposem d'una mostra prou àmplia de la llengua, i per tant podem donar compte de fenòmens discursius de caire general, però sense haver de resoldre qüestions particulars de registres lingüístics molt idiosincràtics. El nostre objectiu final és desenvolupar una eina independent de gènere, però en aquesta primera versió hem reduït el marge de variació possible per centrar-nos en els aspectes més centrals del desenvolupament.

Aquesta mostra és un subconjunt de 14 milions de paraules del corpus periodístic en què hem basat el nostre primer estudi. La unitat de significat és el paràgraf, per tant, recollim les ocurrències d'un marcador en un context no major a un paràgraf. Conté 52.000 oracions en què ocorren un total de 638.788 ocurrències de marcadors (sense comptar signes de puntuació ni estructures sintàctiques)[3]. Els marcadors més freqüents són les conjuncions coordinants, seguits de preposicions i conjuncions subordinants molt gramaticalitzades (vegeu taula 3.1). Els marcadors menys freqüents són, en canvi, locucions adverbials o prepositives, amb un pes lèxic important i un contingut retòric molt explícit: *al fin y a la postre, antes al contrario, con miras a que*, etc. Sovint, es tracta de mots que pertanyen a un registre molt formal o que han caigut en desús. De moltes d'aquestes locucions tan sols en tenim una ocurrència.

En aquest estudi més aprofundit, hem intentat determinar els paràmetres rellevants per la caracterització dels marcadors de cara al seu tractament computacional:

- pel que fa a la seva **funció**
  - quines marques formals en delimiten l'**abast**
  - la seva **relació amb els components textuals** circumdants
  - els **contextos** en què es donen, quins factors contextuals determinen inambiguament la seva funció com a marcadors
- pel que fa al seu **contingut**
  - quins **significats** podem extreure dels marcadors

[3]Hem trobat una ocurrència mitjana de marcadors lèxics en corpus general d'un marcador per oració.

– quina és la **representació** més adequada del significat dels marcadors
– quins **factors contextuals** contribueixen a fer inambigu el significat dels marcadors ambigus

3.2.2 La dimensió delimitadora dels marcadors del discurs

En aquest estudi de corpus, hem trobat que alguns marcadors del discurs assenyalen no tan sols la relació entre segments, sinó que també els delimiten. Així, un marcador pot ser la font d'evidència textual que ens permetrà determinar si ens trobem davant d'un segment o no. Per tant, per realitzar una bona segmentació del text a partir de la qual puguem establir les relacions retòriques entre segments, resulta imprescindible determinar de manera segura i fiable quins són els marcadors que tenen funció delimitadora, és a dir, que ens permeten detectar la presència d'un segment i en marquen els límits. No obstant, alguns marcadors són ambigus en aquest punt: poden tenir una funció oracional, si introdueixen components oracionals, o bé una funció discursiva, si caracteritzen segments retòrics. Vegem aquí un exemple amb el marcador 'mientras':

1. *Simultáneamente a las investigaciones de la fiscalía, el caso de la financiación irregular de la CDU será objeto de una investigación parlamentaria cuyo propósito será determinar si algunas decisiones del Gobierno alemán* ***mientras*** *Kohl fue canciller estuvieron relacionadas con donaciones a su partido.*

2. *Cuando lo hizo, Almaraz sacó la pistola y efectuó seis disparos contra el conductor y su acompañante,* ***mientras*** *Polo le cubría la espalda arma en mano.*

En l'exemple 1, 'mientras' introdueix un segment que forma part del nucli central de l'oració, ja que contribueix a especificar el subjecte, i no en pot ser eliminat sense que canviï el significat de l'element que complementa i, per tant, sense que canviï el significat de l'oració. En canvi, en l'exemple 2 'mientras' introdueix un segment perifèric a l'oració, ja que presenta un event simultani però no lligat intrínsecament al nucli oracional. No hem de confondre aquest fenòmen de doble funció amb la polisèmia que presenten alguns marcadors del discurs, i que per ara no tractarem:

1. ***Mientras*** *el grupo de Greenpeace revisa las costas del Cantábrico, la marea negra asedia las islas de Re y Olerón en su camino hacia el sur.*

2. ***Mientras*** *duren los trabajos, la entrada a la estación de Badal tendrá que hacerse por el acceso de las calles de Carreras Candi y Sants.*

3. ***Mientras*** *no cambie el panorama –con la excepción de algún programa que vale la pena, la mayoría de La 2–, les recomiendo que no enciendan la tele.*

4. *El ajedrecista ruso, que no ha jugado ninguna partida oficial en los últimos seis meses, sigue primero en la clasificación de la Federación Internacional de Ajedrez (FIDE) de primeros de año,* ***mientras*** *el campeón mundial de este organismo, el también ruso Alexander Jalifman, ha subido al 28º puesto.*

> Aquests exemples mostren la polisèmia del marcador 'mientras': en 1 i 2, 'mientras' introdueix segments amb contingut temporal de simultaneïtat, amb la diferència que 1 està orientat al present i 2 al futur, cosa que es marca amb l'indicatiu i el subjuntiu, respectivament. En canvi, en l'exemple 3, 'mientras' introdueix un segment amb significat de condicionalitat, cosa que queda palesa mitjançant el temps subjuntiu i la negació de la frase que l'acompanya. En l'exemple 4, 'mientras' introdueix un segment que contrasta amb la informació anterior.

Com veiem, qüestions com els temps i modes verbals, la polaritat, l'ordre i posició de les frases, etc. són factors que poden contribuir a determinar el contingut dels marcadors ambigus, i potser també la seva funció. Per tal d'explorar aquestes possibilitats, considerem que pot ser útil d'aplicar tècniques d'aprenentatge automàtic sobre un corpus anotat, tal com proposen Passonneau i Litman (1997), que ens permetin donar compte de manera no esbiaixada quines són les claus textuals que permeten la desambiguació dels marcadors ambigus.

Per determinar si un marcador introdueix un segment retòric amb entitat pròpia o un component oracional dependent d'un altre element, hem aplicat una adaptació de la prova que proposen Mann i Thompson (1988) per distingir els segments-nucli dels segments-satèl·lit. Hem aplicat aquesta prova, que no és específica per a l'estudi dels marcadors sinó dels segments perquè considerem que, si podem determinar quan un segment introduït per un marcador és segment o no ho és, podrem determinar si el marcador està fent una funció oracional o discursiva. Així, si partim del nucli verbal d'una oració, hem considerat que un fragment de text introduït per un candidat a marcador del discurs serà un segment retòric si pot ser eliminat de l'oració sense que això comporti:

- ni l'agramaticalitat de l'oració,
- ni tampoc la seva infelicitat[4],
- ni el canvi de significat de l'oració o dels elements que hi resten,
- ni el canvi de significat essencial del discurs

Si es produeix algun d'aquests fenòmens, considerarem que el candidat a marcador del discurs presenta una funció oracional, perquè el fragment de text que introdueix no és un segment retòric sinó un component oracional requerit per algun altre component a nivell sintàctic o semàntic.

exemple: *Pero no está claro* ***si esto es lo que Rusia necesita****.*

> En aquest exemple, el fragment de text introduït per 'si' està requerit pel nucli verbal de l'oració, ja que n'és el subjecte. La seva eliminació suposaria un l'agramaticalitat de l'oració.

exemple: *Que los Reyes Magos nos echen una mano* ***si pueden****.*

> En aquest exemple, el fragment de text introduït per 'si' no està requerit pel nucli verbal de l'oració, sinó que presenta una informació perifèrica. La seva eliminació no suposaria ni l'agramaticalitat de l'oració, ni la seva infelicitat, ni tindria cap repercussió en el seu significat.

Segons alguns experiments (Passonneau i Litman 1997, Marcu 1997) que intenten trobar claus textuals que poden contribuir a determinar la funció dels marcadors ambigus, semblen especialment reveladores les pauses, en el discurs parlat i, en els textos escrits, els seus correlats ortogràfics. Podríem dir que les pauses, ja siguin d'entonació o de puntuació, contribueixen a delimitar la part central de l'oració, on ocorren els components requerits per a la seva bona formació sintàctica i semàntica. Fora d'aquesta columna vertebral de l'oració podem trobar elements perifèrics que tenen un lligam diferent, sovint més laxe, amb el nucli oracional. És en aquesta posició perifèrica que podem considerar que els marcadors del discurs amb ambigüitat funcional presenten de manera sistemàtica una funció discursiva.

exemple: *El Gobierno de José María Aznar optó* ***por asumir como un gran mérito la falta de incidentes relevantes.***

> En aquest exemple, el nucli oracional inclou el component introduït per la preposició 'por', que està requerit pel verb 'optar'

[4]Com ja hem dit, recorrem al concepte d'infelicitat, més vague i intuïtiu que el d'agramaticalitat, perquè no disposem, en el nivell discursiu, d'unes regles de bona formació que expliquin de manera determinant per què una seqüència discursiva mai no seria produïda, en circumstàncies normals, per un parlant competent d'una llengua.

exemple: *La gripe, de 48 horas de incubación, es la inflamación de todas las mucosas de las vías respiratorias altas,* ***por ello se puede evitar el contagio persona a persona usando una sencilla mascarilla de papel****, como se ha practicado siempre en Oriente.*

> En aquest exemple, el nucli oracional estaria format per '*La gripe es la inflamación de todas las mucosas de las vías respiratorias altas',* i la resta d'elements de la oració són perifèrics, com bé ho evidencien les pauses sintàctiques.

No obstant, hem de tenir en compte fenòmens sintàctics que poden crear soroll en aquesta norma que acabem de constatar, com les dislocacions de components que pertanyen al nucli oracional:

exemple: *Que puedan llegar a inspirar compasión a todo un ministro del Interior, ha de ser trágico para ellos.*

> En aquest exemple trobem una oració no normativa, ja que apareix una coma entre el subjecte i el predicat, de manera que, sense informació sintàctica, podríem determinar, equivocadament, que el nucli oracional és el que no té cap marca de subordinació, i el que està separat per comes d'aquest nucli està en posició perifèrica.

Observant aquestes i altres qüestions que afecten la funció dels candidats a marcadors del discurs ens adonem que la informació sintàctica i la informació retòrica estan estretament imbricades. Efectivament, és gràcies a informació de gramàtica oracional que podem determinar la funció dels marcadors ambigus: podem dir que, si un ítem lèxic introdueix un fragment de text que pertany al nucli central d'una oració, aquest ítem lèxic no té funció de marcador del discurs, sinó que té una funció d'àmbit estrictament oracional. Podem determinar que un component oracional forma part d'aquest nucli oracional si:

– està dominat per algun dels components oracionals que formen part d'aquest nucli

– és un component imprescindible de la configuració prototípica de l'escenari de l'event

Tenint en compte aquestes qüestions, hem classificat els marcadors que hem estudiat en marcadors amb funció discursiva inambigua i marcadors amb doble funció. Els marcadors de tipus contraargumentatiu (*aunque, a pesar de,...*) solen presentar una funció inambiguament discursiva, ja que la natura del seu significat fa que no puguin estar inclosos en el nucli de cap event. També hem vist que els marcadors amb un pes lèxic important en la forma, és a dir, poc gramaticalitzats (*con la condición de, como consecuencia de,...*), solen tenir sistemàticament una funció discursiva: probablement, perquè el propi pes lèxic adquireix un estatus de límit de nucli oracional, amb la mateixa força que poden tenir els signes de puntuació.

Pel que fa als marcadors amb doble funció, hem trobat:

1. peces lèxiques que poden estar **subcategoritzades** per algun component oracional, sobretot els nuclis verbals: bàsicament, es tracta de conjuncions subordinants o preposicions. Fins que no puguem saber de manera inambigua quan un predicat verbal requereix algun d'aquests elements o no, és a dir, fins que no puguem determinar la valència d'un verb, no podrem determinar si una instància donada d'un marcador d'aquest tipus fa una funció oracional o discursiva.

2. components que solen formar part intrínseca de la **configuració prototípica de l'event**, i, per tant, solen formar part del nucli oracional

   (a) components de tipus circumstancial de temps, lloc, manera, ja siguin sintagmes adverbials, sintagmes nominals o frases subordinades

   (b) components que expliciten la relació causal entre dos events, que podem considerar com un únic event per la seva estreta relació

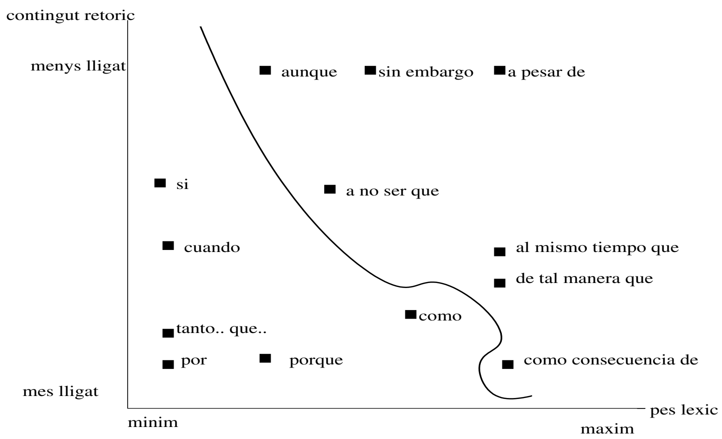

Figura 3.1: Simulació de càlcul de la funció delimitadora dels marcadors en un espai bidimensional

Així doncs, el fet que un marcador pugui tenir doble funció sembla que està determinat per dos paràmetres: per una banda, el contingut retòric del marcador, és a dir, si el marcador introdueix un component oracional que, tal com acabem d'exposar, estigui fortament lligat a la configuració prototípica de l'event pel seu significat. Per altra banda, el pes lèxic del marcador sembla un factor decisiu pel que fa a la funció discursiva. Així, els marcadors amb un pes lèxic major solen tenir una funció discursiva més clara que els marcadors més gramaticalitzats. Sembla que aquests dos paràmetres contribueixen de manera col·laborativa per determinar la funció de cada marcador determinat. No obstant, fins ara tan sols hem recollit intuïcions i tendències. Esperem que podrem arribar a conclusions més objectives a partir d'un estudi sobre corpus dels marcadors del discurs en què, aplicant tècniques d'aprenentatge automàtic, puguem sistematitzar les interaccions entre els diferents paràmetres definitoris dels marcadors (vegeu figura 3.1).

En la implementació de l'eina de segmentació retòrica automàtica dels textos, hem tractat de manera diferent els marcadors segons la seva ambigüitat com a delimitadors. Així, els marcadors amb doble funció tan sols identifiquen segment de manera inambigua en un context de pausa sintàctica que evidenciï la seva posició perifèrica respecte del nucli central de l'oració, és a dir, separats d'aquest nucli per una coma. Quan no es dóna aquest context, hem optat per identificar-los com a segments possibles, que hauran de ser tractats amb més profunditat i més coneixement en un procés d'interpretació retòrica posterior. Aquest procés incorporarà alhora el coneixement retòric i el coneixement sintàctic i discursiu aprofundit que resulta imprescindible per una bona anàlisi, com hem mostrat en aquest apartat.

3.2.3 Descripció de marcadors del discurs

A partir de les concluscions a què hem arribat en l'estudi de corpus, i amb l'objectiu de desenvolupar una eina que permeti l'anàlisi retòrica automàtica dels textos, hem elaborat una descripció sistemàtica i formal d'un subconjunt de 388 marcadors que hem estudiat (vegeu Annex IV). Hem establert els següents paràmetres no excloents:

1. **funció relacional**: segons la informació que es pot extreure de cada marcador i la fiabilitat d'aquesta informació, els descrivim com a relacionadors o delimitadors. Aquestes dues classes no són disjuntes (vegeu figura), sinó que trobem marcadors que poden informar alhora sobre els límits dels segments en què es troben i sobre la relació entre aquests segments. La resta de paràmetres que exposem s'apliquen només als relacionadors, ja que descriuen aspectes dels marcadors relacionats exclusivament amb la informació de relació.

(a) **delimitadors** (213 marcadors)[5]: marques de límits de segment. Els delimitadors prototípics són els signes de puntuació, especialment les cometes i els signes parentètics, que no aporten cap tipus d'informació retòrica. També tractem únicament com a delimitadors els punts i les conjuncions coordinants, ja que no tenen cap contingut propi més que el d'addició de text. No obstant, en context, aquestes darreres marques poden omplir-se amb contingut retòric. L'exemple més sistemàtic és el de la relació de llista, en què una seqüència de comes i conjunció coordinant tenen un contingut retòric de distribució de la informació, encara que aquests elements poden adquirir gairebé qualsevol contingut retòric que impliqui seqüència.

exemple: *La amenaza del efecto 2000 no se consumó* ***y*** *el planeta pudo celebrar su mayor fiesta sin sobresaltos* ***ni*** *caos en los servicios considerados estratégicos.*

En aquest exemple, 'y' funciona com a delimitador dels segments a la seva dreta i esquerra, però 'ni' coordina elements inferiors al segment i, per tant, no té estatus de delimitador.

exemple: *un violador nunca será* ***mister, maestro o profesor, ni un secuestrador, torturador o ladrón*** *recibirá el trato de cortesía que supone anteponer a su nombre la voz don.*

En aquest exemple, les comes i les conjuncions haurien de ser considerades com a marcadors de llista, mentre que 'ni' delimita dos segments diferenciats.

Com hem dit en l'apartat anterior, alguns marcadors del discurs són ambigus en aquest paràmetre, i així ho reflectim en la descripció: aquells que tenen una funció delimitadora inambigua es caracteritzen com a delimitadors '**forts**'(106 marcadors), mentre que els que presenten doble funció es caracteritzen com a delimitadors '**febles**'(107).

(a) **relacionadors** (173 marcadors): marques explícites de relacions retòriques.

exemple: *La mayoría de los moscovitas acogieron la dimisión de Yeltsin como un regalo de año nuevo,* ***aunque*** *muchos de ellos escucharon ante el televisor con lágrimas en los ojos sus peticiones de perdón.*

En aquest exemple, el marcador 'aunque' determina que la frase subordinada és un satèl·lit de la frase principal i expressa la relació retòrica de concessió.

Un marcador del discurs pot tenir, alhora, funció relacionadora i delimitadora. No obstant, en l'estat actual de la nostra investigació, tan sols hem pogut donar compte de la funció delimitadora dels marcadors de tipus sintàctics incomplets (conjuncions i preposicions). La distribució dels marcadors de tipus sintàctics complets no segueix patrons clars, per tant, la seva funció com a límits de segment no és sistemàtica i no podem utilitzar-la. Una possible via per donar compte del comportament d'aquests marcadors és aplicant tècniques d'aprenentatge automàtic sobre un corpus anotat. Intentarem dur a terme un experiment per obtenir aquesta informació, de manera que puguem utilitzar també aquestes marques per a delimitar segmens.

2. **funcionament sintàctic** dels relacionadors: no hem especificat el funcionament sintàctic dels delimitadors perquè és simple: es correspon amb el d'una conjunció coordinant. Pel que fa als relacionadors, distingim:

   (a) relacionadors **sintàcticament incomplets** (211 marcadors): van indefectiblement units a un component sintàctic: una frase subordinada, un grup nominal, etc. Distingim:

[5]Els nombres de marcadors es donen sobre la mostra no expandida de 388 marcadors de l'Annex III, però el nostre lexicó de treball compta amb 610 formes, resultat d'expansions regulars de les formes bàsiques.

i. **subordinant** (106): comportament sintàctic típic de conjunció subordinant. Com a delimitador, separa la frase principal i la subordinada com a segments diferents, i com a marcador estableix una relació retòrica entre elles. La frase subordinada sol establir una relació de satèl·lit amb la principal, que sol ser el nucli.

exemple*:** *Una hora y media después,* ***cuando *el avión entraba en el espacio aéreo indio, los secuestradores se dieron a conocer.*

En aquest exemple, el marcador 'cuando' determina que la frase subordinada és un satèl·lit de la frase principal y expressa la relació retòrica de simultaneïtat en el temps.

exemple*:** *El déficit es mayor en las donaciones de sangre,* ***a pesar de que *el índice aumentó un 2% en 1999.*

En aquest exemple, el marcador 'a pesar de' es combina amb la conjunció subordinant 'que' per establir una relació concessiva entre dos segments de text.

***exemple*:** *las cámaras de TVE permanecieron en la Zarzuela hasta una hora antes de la emisión del discurso, la noche del día 24,* **por si** *se producía una acción de ETA que obligara a cambiar el mensaje grabado.*

En aquest exemple, el marcador 'por' es combina amb la conjunció 'si' per establir una relació que suma els continguts d'ambdues (causal-condicional).

ii. **preposicional** (94): tot element que requereix un sintagma nominal o similar per a ser sintàcticament complet.

exemple*:** *En el 2002 comenzará el periodo de espera, que durará 25 años,* ***con objeto de *que se reduzcan los niveles de radiactividad en el interior del edificio del reactor.*

En aquest exemple, el marcador 'con objeto de' indica que el segment que introdueix, la frase subordinada, és el satèl·lit d'una relació de finalitat.

exemple*:** *Los médicos asturianos consideraron que el cuerpo de la madre sería una incubadora ideal,* ***a pesar de no poder facilitar al feto los estímulos habituales en una gestación normal*.*

En aquest exemple, la combinació d'un infinitiu amb un element prepositiu explicita una relació concessiva entre la frase d'infinitiu i la frase principal.

iii. **conjunció** (11): elements amb el comportament de conjunció coordinant (adversativa, disjuntiva)

***exemple*:** *Podría hablarse de tragicomedia* **pero** *Echanove lo define como "un esperpento moderno".*

En aquest exemple, el marcador 'pero' indica que els dos segments que relaciona tenen una relació adversativa entre ells, però no explicita si algun d'ellos és satèl·lit de l'altre.

iv. **discontinus**: marcadors formats per dos elements sintàcticament diferenciats però que expressen una relació retòrica única. Per la seva complexitat, encara no podem tractar-los.

exemple*:** *"No se imaginan los americanos el balazo que les va a dar", comentó Hayek,* ***tan *amiga de Penélope Cruz* ***que*** *la tiene hospedada en casa.*

En aquest exemple, els marcadors 'tan' i 'que' expressen una única relació consecutiva.

(b) **relacionadors sintàcticament complets** (173 marcadors): no requereixen cap element per a la seva bona formació sintàctica, sinó que són autònoms. Tenen un comportament distributiu molt divers, del qual encara no hem pogut donar compte.

i. **adverbial** (136): té el comportament sintàctic que podríem atribuir a un complement circumstancial, es tracta sobretot de sintagmes preposicionals o adverbials.

exemple*:** *Por ahora, sus gestos se reducen a la intimidad del vestuario. Ayer,* ***por ejemplo*, sorprendió a todos (técnicos, jugadores y auxiliares) con un obsequio conmemorativo del Balón de Oro que le otorgó la revista France Football hace un par de semanas.*

En aquest exemple, el marcador 'por ejemplo' determina que el segment en què es troba és un satèl·lit del segment precedent.

ii. relacionadors composicionals **anafòrics** (37): combinacions de marcadors amb un element anafòric.

exemple*:** *El Dakar me apasiona desde hace 20 años, pero por encima de esa pasión, amo Africa y especialmente a su gente,* ***por eso *me entristece tanto una situación como ésta.*

En aquest exemple, el marcador 'por eso' relaciona el segment a la seva dreta com un satèl·lit d'una relació de causa amb el segment de la seva esquerra.

iii. relacionadors composicionals **catafòrics**: combinacions de marcadors amb un element catafòric. Aquest tipus de marcadors no és freqüent, per la qual cosa hem contemplat la seva existència però no hi hem aprofundit.

(c) **semirelacionadors**: elements que constitueixen alhora el marcador d'una relació retòrica i un dels segments implicats en la relació.

i. **no personals**: frases amb el verb principal en gerundi o participi, en una posició sintàctica que evidencia una relació retòrica (entre pauses)

exemple*:** ***Solucionados los problemas de financiación*, Nadal explicó que los retos de este año de su gobierno se concretarán en promover más vínculos entre los estudios de bachillerato y la universidad.*

En aquest exemple, la frase de participi expressa una relació de circumstància a la frase principal per la seva posició sintàctica, separada de la resta de l'oración per una coma.

ii. **relatius**: en determinades posicions sintàctiques (entre comes), els pronoms relatius evidencien una relació retòrica d'excurs.

exemple*:** *Una vieja norma periodística,* ***que no sé si ha sido abolida de los manuales*, dice que los criminales han de ser tratados sin consideración social o de rango alguna.*

En aquest exemple, la frase de relatiu entre comes està en relació d'excurs amb la frase principal.

iii. **adverbials**: en posició interpausal (entre comes o punt i/o coma), un grup adverbial evidencia el seu estatus de segment no directament vinculat al nucli oracional, i, per tant, podem considerar que estableix una relació de tipus textual amb aquest nucli.

iv. **preposicionals**: en posició interpausal (entre comes o punt i/o coma), un grup preposicional evidencia el seu estatus de segment no directament vinculat al nucli oracional, i, per tant, podem considerar que estableix una relació de tipus textual amb aquest nucli.

v. **aposició**: en posició interpausal (entre comes o punt i/o coma) i en una certa posició dins de l'oració, un nom propi evidencia el seu estatus de segment no directament vinculat al nucli oracional, i, per tant, podem considerar que estableix una relació de tipus textual amb aquest nucli.

vi. **adjectius**: en posició interpausal (entre comes o punt i/o coma), un grup adjectiu evidencia el seu estatus de segment no directament vinculat al nucli oracional, i, per tant, podem considerar que estableix una relació de tipus textual amb aquest nucli.

3. **comportament retòric**: pel que fa a la distribució de la rellevància discursiva que expliciten els marcadors, hem codificat cinc tipus generals de comportatment. Val a dir que, ara com

ara, només donem compte de les relacions discursives de curt abast, per tant, tan sols hem estudiat la relació dels relacionadors amb els segments circumdants.

(a) **satel·litzador** (250 marcadors): explicita que un segment de text és el satèl·lit d'una relació en què podem distingir nucli i satèl·lit.

exemple: *ACCIONA salió de mínimos anuales con bastante fuerza. Las acciones, **no obstante**, se encuentran un 35% por debajo del cierre del pasado año.*

En aquest exemple, el segment en què es troba inclòs el marcador 'no obstante' és el satèl·lit d'una relació de concessió, mentre que el segment precedent n'és el nucli.

exemple: *El miércoles llega el Oporto al Camp Nou y, a partir de entonces, no habrá respiro. "En marzo, todos los partidos serán decisivos", añadió, **sin olvidar que** el Madrid está en la misma situación: sigue vivo en las tres competiciones.*

En aquest exemple, el segment que introdueix 'sin olvidar que' és el satèl·lit d'una relació de digressió, de la qual el segment anterior és el nucli.

(b) **nucleitzador** (44 marcadors): explicita que un segment de text és nucli d'una relació en què podem distingir nucli i satèl·lit.

exemple: *Sanitat atribuye el aumento de defunciones a la llegada de la gripe. Los máximos índices de casos de gripe, **no obstante**, se producirán durante enero y febrero próximos, a medida que surjan las olas de frío, advierten los técnicos de Sanitat.*

En aquest exemple, el segment en què es troba 'no obstante' és el nucli d'una relació de concessió, de la qual el segment anterior és el satèl·lit.

exemple: *Se trata de una de las últimas zonas vírgenes del Pirineu, donde viven especies tan escasas como el oso pardo, el urogallo, la perdiz blanca o la nutria. También cuenta con grandes densidades de mamíferos: rebecos, ciervos, corzos, etcétera. Es, **en definitiva**, un paraíso natural.*

En aquest exemple, el segment en què es troba 'en definitiva' és el nucli d'una relació de concessió, de la qual els dos segments anteriors són el satèl·lit.

(c) **organitzador** (50): explicita la distribució de la informació en el text, sense que hi hagi cap relació de nucli i satèl·lit entre els segments implicats. Distingim:

i. **introductor** (15): explicita la introducció d'informació (*el hecho es que*, *el caso es que*, etc.).

exemple: *El silencio de González es más notorio desde que el presidente de Castilla-La Mancha,José Bono, presentó su candidatura, el pasado sábado, y afirmó saber que el ex líder socialista no quiere presidir el partido. **El hecho es que** Bono no desea que un hipotético presidente del PSOE pueda proyectar sombra sobre su liderazgo.*

En aquest exemple, el segment introduït per 'el hecho es que' aporta informació al text, però no podem determinar la major o menor importància d'aquesta informació respecte de la informació anterior.

ii. **distribuidor** (12): distribueix la informació en dos o més segments diferenciats alhora que relacionats entre ells mitjançant un o més marcadors. La relació entre els segments és d'organització argumentativa (*por una parte / por otra parte, por un lado / por otro lado*, etc.).

exemple: *El debate sobre el desfile de las fuerzas armadas partió ayer el Ayuntamiento **en dos: por un lado**, PSC y PP, tradicionales adversarios, que se aliaron para rechazar las mociones en contra de la parada militar; **por otro lado,** CiU, ERC e IC, que de manera individual coincidieron en oponerse a la celebración.*

En aquest exemple, 'por un lado' i 'por otro lado' distribueixen la informació entre els segments que introdueixen respectivament i els posen en el mateix nivell de rellevància discursiva. Podem considerar que 'en dos' com una clau que contribueix a determinar la natura d'aquests marcadors bipartits.

iii. **seqüenciador** (18): distribueix la informació en dos o més segments diferenciats alhora que relacionats entre ells mitjançant un o més marcadors. La relació entre els segments és de seqüència argumentativa (*en primer lugar / en segundo lugar /..., primero / segundo /...*, etc.).

***exemple*:** *Efectivamente, este puede ser el paso más importante de la integración económica y financiera europea desde la implantación del euro, por **un triple motivo**. **En primer lugar**, porque, al unir a los dos primeros mercados del Viejo Continente, fuerza al resto de ellos a seguir esa pauta, constituyendo una bolsa verdaderamente paneuropea. **En segundo lugar**, porque sitúa al Reino Unido en una tesitura -las acciones del nuevo mercado se cotizarán en euros- que sólo puede desembocar razonablemente en la adhesión de ese país a la unión económica y monetaria (UEM). Y, **en tercer lugar**, porque, al establecer un acuerdo con el mercado norteamericano de acciones de empresas de alto crecimiento o nuevas tecnologías (Nasdaq), facilita el objetivo establecido por la Unión Europea en la cumbre de Lisboa de apostar a este lado del Atlántico por las tecnologías de la información como mejor garantía del pleno empleo.*

En aquest exemple, 'en primer lugar', 'en segundo lugar' i 'en tercer lugar' distribueixen la informació dels segments que introdueixen, la posen en un mateix nivell de rellevància discursiva i l'ordenen de manera seqüencial. Podem considerar 'un triple motivo' com una clau que contribueix a determinar la natura d'aquests marcadors tripartits.

(d) **connector** (43): evidencia una relació retòrica entre dos segments que no pot definir-se en termes de nucli-satèl·lit ni de distribució de la informació.

***exemple*:** *Sólo que aquí se trataba de una etiqueta de sólo 7 por 2 centímetros, **es decir**, prácticamente invisible.*

En aquest exemple, el marcador 'es decir' relaciona dos segments de manera que el segment a la seva dreta és un comentari sobre el segment a la seva esquerra.

***exemple*:** *Un año antes, el 31 de diciembre de 1998, de los 177.843 millones de euros a que ascendía el saldo de bonos y obligaciones, 40.843 millones estaban en manos de inversores foráneos, **es decir**, el 23%.*

En aquest exemple, el marcador 'es decir' relaciona dos segments de manera que el segment a la seva dreta resumeix la informació que aporta el segment a la seva esquerra. A través d'aquest exemple i l'anterior veiem com resulta impossible determinar mitjançant característiques sistemàtiques i formals quin dels dos segments és el més rellevant, però tampoc no podem parlar de distribució de la informació.

4. **comportament segmental**: hem sistematitzat els esquemes de relació entre els marcadors i els segments que els envolten, tot i que, en el cas dels marcadors de tipus sintàctic complet, l'esquema tan sols reflecteix la tendència majoritària, ja que presenten una gran llibertat de posició que fa gairebé impossible donar compte de manera segura del seu context d'aparició. Quan un marcador explicita una relació retòrica amb nucli i satèl·lit, considerem que el marcador està més estretament lligat amb el segment satèl·lit. Hi ha una estreta relació entre el comportament segmental, retòric i sintàctic dels marcadors:

   (a) **a la dreta** (243 marcadors): el segment més estretament lligat al marcador és a la seva dreta. Aquest paràmetre és redundant pels relacionadors sintàcticament incomplets, que tenen sempre un comportament segmental a la dreta, però és rellevant pels complets: un 24% dels marcadors adverbials semblen estar més lligats al segment de la seva dreta.

exemple: *El milagro ha llegado con la aportación de 1.584 millones de la Generalitat,* ***como resultado de*** *un convenio suscrito entre el consistorio y el gobierno catalán.*

(b) **a l'esquerra** (54 marcadors): el segment més estretament lligat al marcador és a l'esquerra. S'aplica només a marcadors sintàcticament complets (un 29% dels complets estan més lligats al segment de l'esquerra), preferentment anafòrics (28 anafòrics i 22 adverbials)

exemple: *Craxi se defendió hasta hace pocos días de las acusaciones de corrupción, aduciendo que en Italia "todos los partidos alimentaban sus arcas con el mismo sistema", y que,* ***por esta razón***, *él no era más culpable que los demás. Por eso pedía una comisión investigadora y la pide el centroderecha liderado por su amigo Silvio Berlusconi.*

(c) **ambigu a dreta o esquerra** (3 marcadors): un dels segments està més estretament lligat al marcador, però no disposem de marques formals per determinar quin. S'aplica als marcadors de tipus adverbial i anafòric.

exemple: *En algunas áreas, en especial aquellas más hundidas respecto al nivel del mar, el arroz es el único cultivo posible porque los terrenos inundados evitan la salinización.* ***Así***, *si las explotaciones no son rentables, no hay alternativa posible.*

exemple: *Afortunadamente, en algunas ocasiones la línea de lo que parece irremediable se quiebra.* ***Así***, *hace apenas tres años, el viejo camino fue reabierto y se desbrozó y limpió con el esfuerzo voluntario de muchos vecinos de Vandellòs y de l'Hospitalet de l'Infant.*

(d) **per inclusió** (69 marcadors): el segment més estretament lligat al marcador és aquell en què es troba sintàcticament inclòs. S'aplica només a marcadors sintàcticament complets (un 39% dels marcadors complets).

exemple: *El balance de soldados israelíes muertos se eleva a mil y el de heridos, a miles. El criterio favorable podría cambiar,* ***sin embargo***, *si los territorios fronterizos con Líbano pagaran la retirada viéndose obligados a mantenerse en perpetuo estado de alarma y bajo el fuego de las armas de Hezbollah.*

(e) **bidireccional** (18 marcadors): cap segment està més estretament lligat al marcador. S'aplica a organitzadors i connectors de tipus adverbial i conjunció.

exemple: *Y en segundo lugar, que sufre una neuropatía diabética,* ***esto es***, *un transtorno de los nervios de las extremidades*

5. **contingut retòric**: basant-nos una tipologia de relacions retòriques operativa, hem associat a cada marcador un contingut retòric concret. Trobem que una tipologia de relacions retòriques explicita de manera adequada el coneixement lingüístic associat als marcadors. per exemple, hem observat comportaments comuns en alguns d'aquests grups de significat:

   (a) les qüestions d'interrupció del fluxe informatiu solen seguir els mateixos patrons en tots els segments marcats amb relacions retòriques de tipus circumstancial.

   (b) els elements que expressen condició solen establir una relació sistemàtica de nucli-satèl·lit entre la pròtasi (nucli) i l'apòdosi (satèl·lit).

   (c) els marcadors adversatius, si són del tipus 'subordinant', tenen una distribució sistemàtica de nucli-satèl·lit.

   (d) els marcadors de causalitat presenten una fenomenologia comuna, encara que el grup que la manifesta més acusadament és el de les relacions consecutives. En aquest tipus de relacions resulta molt difícil d'establir una relació nucli-satèl·lit útil per a resum automàtic.

Així doncs, un marcador podrà estar caracteritzat amb tots els atributs que hem descrit. Si es tracta d'un relacionador, haurà de poder ser descrit de manera inambigua en cadascun d'aquests paràmetres per al seu ús efectiu en el procés d'interpretació retòrica d'un text. Si no es pot obtenir tota la informació requerida d'un relacionador, o no és prou exhaustiva o inambigua, tan sols s'utilitzarà la informació de delimitació que aporti. Tot seguit aportem alguns exemples de marcadors del discurs analitzats segons els paràmetres que acabem d'exposar:

exemple*:** *Stallone se lo ha tomado con paciencia* ***y *ya se ha resignado,* ***aunque*** *intenta hacer mínimas variaciones.*

> En aquest exemple, **y** és un delimitador, ja que tan sols aporta informació inambigua pel que fa al límit de segment. Podríem considerar que, en aquest cas, està fent la funció de relacionador del tipus conjuntiu, connector bidireccional i coordinador, però no disposem de marques formals que ens permetin determinar-ho de manera automàtica. D'altra banda, **aunque** és un relacionador inambigu, sintàcticament incomplet ja que és un subordinant, satel·litzador a la dreta com la major part de subordinants, i amb un significat retòric de concessió.

***exemple*:** *Podría hablarse de tragicomedia* **pero** *Echanove lo define como "un esperpento moderno".*

> En aquest exemple, el marcador **pero** és un relacionador que explicita una relació de concessió entre els dos segments que formen l'oració. Es tracta d'un relacionador de conjunció, un connector bidireccional, ja que no podem considerar que cap dels segments sigui més nuclear que l'altre en la relació.

3.3 Conclusions

Hem presentat una primera aproximació al tractament computacional de les paraules clau d'estructura textual, els anomenats marcadors del discurs. En primer lloc, hem determinat quin tipus de marques textuals ens podien ser útils per al nostre objectiu aplicat, l'obtenció d'una estructura retòrica del text. Aquesta estructura ha de donar compte de les parts de què es composa un text. En un segon nivell, aquesta estructura ha de reflectir la importància relativa de cadascuna d'aquestes parts en el conjunt del text. Aquesta importància estarà determinada per les relacions entre els segments de text que expliciten els marcadors del discurs. Així doncs, hem proposat una definició de marcador del discurs que és útil per als nostres objectius.

Hem dut a terme un estudi de marcadors del discurs en espanyol, i com a resultat, hem recollit un primer conjunt de marcadors que hem sistematitzat en un lexicó (Annex IV). Per fer-ho, hem detectat els paràmetres descriptius necessaris per a l'ús computacional d'aquestes marques i hem descrit cada marcador seguint aquests paràmetres. Un cop haguem comprovat l'efectivitat d'aquesta primera mostra de marcadors del discurs pel reconeixement de l'estructura retòrica d'un text, se'ns obre la possibilitat d'incorporar nous candidats a marcadors del discurs. Aquests candidats seguiran un procés d'estudi similar al de la mostra amb què treballem ara com ara. Algunes fonts possibles per la identificació de candidats a marcadors del discurs són:

- fonts lexicogràfiques
- estudi de corpus
- ampliació de les relacions discursives que tractem
- mecanismes lèxics i sintàctics de composició de marcadors del discurs, com per exemple:
  - sinònims o cuasi-sinònims dels marcadors ja establerts, que podem detectar mitjançant la substitució paradigmàtica dels diversos ítems

  - extensions morfològiques d'arrels lèxiques amb contingut retòric explícit: p. ex., podem estudiar i sistematitzar els derivats de *concretar* (*concretando*, *concretamente*), *especificar*, *resumir*, etc., així com les seves ocurrèncie sístemàtiques com a nuclis d'estructures sintàctiques (*en resumen*, *más concretamente*).
  - estructures sintàctiques productives com a marcadors: construcciones com la de sintagma preposicional amb elements anafòrics: *por eso, a pesar de todo*, etc.

- les cadenes de paraules més freqüents després de signes de puntuació: en aquestes posicions solen ocórrer amb molta freqüència els marcadors del discurs. També podem detectar altres posicions d'ocurrència de segments freqüents i extreure'n la informació.

- elements dels quals es pot deduir la relació retòrica amb els elements circumdants pel seu context sintàctic: per exemple, sintagmes nominals, adverbials o preposicionals entre pauses:

  exemple: *El ministro de Interior, Mariano Rajoy, declaró ayer...*

  > El sintagma nominal entre comes està en relació d'aposició amb el sintagma nominal precedent.

  exemple: *La ministra de Cultura, en la foto de la derecha, dejó claro que...*

  > El sintagma preposicional entre comes expressa una circumstància gens central a l'event de l'oració

- marcadors dependents de domini per aplicacions concretes: creiem que en dominis específics
  - poden haver-hi marcadors també específics,
  - alguns marcadors de significat molt general poden tenir un significat més concret
  - es pot reduir l'ambigüitat dels marcadors

D'altra banda, hem trobat que la interacció entre la informació retòrica i la informació sintàctica dels textos és altament informativa. Si bé és cert que en el processament superficial de l'estructura del discurs s'ha treballat majoritàriament a partir de marques lèxiques, com són els marcadors del discurs (Edmunson 1969, Passoneau i Litman 1993, Ono 1994, Marcu 1997), trobem que la informació morfosintàctica, encara que sigui parcial, pot aportar una informació molt valuosa al tractament d'aquestes marques, perquè contribueix a:

- determinar quines formes ambigües són efectivament marcadors del discurs.
- delimitar l'abast dels marcadors.
- identificar alguns marcadors que ho són tan sols per la seva funció sintàctica, els que anomenem 'semirelacionadors': frases no personals, frases de relatiu, i també algunes construccions sintàctiques en determinats contextos: sintagmes preposicionals, adverbials o adjectius en posició interpausal (entre comes o punts i/o comes).

Però no és tan sols la sintaxi que contribueix a la retòrica, sinó que també el procés invers resulta informatiu: així, els límits de segment poden contribuir de manera decisiva a la delimitació de les estructures sintàctiques. En un nivell més profund, la imbricació entre sintaxi i retòrica permetria determinar quins components d'una oració són subcategoritzats pel predicat i quins, en canvi, hi fan una contribució de caire retòric. Aquestes intuïcions semblen mostrar com una aproximació coordinada a un mateix objecte, l'oració, des de diferents nivells lingüístics, pot resultar enriquidora en tots els aspectes.

Capítol 4

Segmentació

Hem considerat que el primer pas per donar compte de l'estructura retòrica d'un text és la segmentació del text en unitats discursives mínimes, els anomenats segments. Per fer-ho, caldrà, en primer lloc, determinar què és un segment: tot seguit presentem una primera definició que pot ser útil per al nostre objectiu alhora que intenta constituir una base formal sòlida. També presentarem algunes qüestions de relacions entre segments que afecten directament la segmentació. Donat que volem treballar amb unitats de nivell oracional i intraoracional, adquireixen especial importància les relacions entre sintaxi i segmentació, estretament lligades amb les relacions entre sintaxi i marcadors del discurs (apartat 3.2.3).

En segon lloc, presentem el disseny, implementació i avaluació d'una eina de segmentació retòrica que porta a la pràctica les qüestions teòriques que hem treballat fins aquest moment. Malgrat que aquesta eina ha estat concebuda com un mòdul dins d'un sistema més ampli de processament del llenguatge natural, creiem que també pot ser útil en altres aplicacions.

4.1 Definició de 'segment'

La segmentació sol constituir el pas previ ineludible per a l'anàlisi dels textos en PLN. Ara bé, cada línia de treball defineix com a 'segment' el tipus d'unitat lingüística més adient a les necessitats que ha de cobrir: trobem conceptes diferents de 'segment' en anàlisi morfològica, en traducció automàtica (unitats de traducció), en alineació de corpus paral·lels, en extracció d'informació, en classificació de documents, en recuperació d'informació, etc. Aquestes unitats de treball van des de les cadenes de caràcters entre dos punts (EAGLES 2000) fins a conjunts de paràgrafs que desenvolupen un mateix tema (classificació de documents, RI), passant per unitats multiparaula (TD, alineació de corpus paral·lels), cadenes de paraules entre dos punts (anàlisi sintàctica, alineació de corpus, EI) i fins i tot parts d'un document determinades pel seu format (TA, alineació de corpus (Martínez 1999)). Nosaltres ens centrarem en la unitat segment en el nivell discursiu.

Tradicionalment, la lingüística s'ha centrat en la paraula i l'oració com a unitats de significat, deixant de banda el nivell textual. És per això que manca una tradició teòrica sòlida i coherent pel concepte de segment discursiu. Però a partir dels anys 60, des de la pragmàtica i la lingüística textual es va començar a considerar el text com a quelcom més que una seqüència d'oracions. És aleshores que s'identifiquen algunes característiques exclusivament textuals, com ara la cohesió (Halliday i Hasan 1976), coherència, informativitat, intencionalitat, situacionalitat, intertextualitat, etc. (Dressler i Beaugrande 1997). Aquests trets defineixen el text com a objecte lingüístic i comunicatiu qualitativament diferent de la resta. Alhora, són paràmetres vàlids per donar compte de la seva estructura interna, tant pel que fa a les parts que el conformen com a les relacions que s'estableixen entre aquestes parts.

Pel que fa a la lingüística computacional, els segments de nivell textual s'han treballat en les àrees que requereixen la comprensió profunda del llenguatge natural, especialment TA, EI, generació automàtica d'informes, resum automàtic, etc. En aquestes àrees s'exploten especialment les característiques més estructurals dels textos, molt informatives respecte de la natura textual. També és cert que la informació de tipus estructural és molt adequada als objectius i limitacions de la intel·ligència artificial, ja que es resulta molt formalitzable i, per tant, la seva implementació en sistemes informàtics resulta factible.

Des de la perspectiva computacional s'han treballat diverses aproximacions a l'estructura textual. S'han proposat diversos tipus de relacions possibles entre els elements que formen un text, tenint en compte els diversos tipus d'informació textualment rellevant de què hom pot donar compte computacionalment: segons el contingut semàntic (Polanyi 1995), les intencions comunicatives (Carberry 1990, Grosz i Sidner 1986), l'estructura atencional (Grosz i Sidner 1986), les relacions de coherència (Hobbs 1985, Mann i Thompson 1988), etc. Ara bé: encara resta per resoldre de manera científicament rigorosa i satisfactòria com es poden determinar els elements entre els quals s'estableixen aquestes relacions. En general, s'adopta una postura operativa, de manera que la definició de segment es caracteritza per la seva utilitat per cada treball particular.

Creiem que aquesta postura és certament útil per al procés de la investigació sobre estructura textual. Nogensmenys, a mesura que es construeixen teories i treballs, es va posant de manifest una necessitat creixent per una definició sòlida i consensuada d'aquest concepte. Tot seguit presentem algunes reflexions sobre aquesta qüestió. No obstant, no hem arribat a cap conclusió definitiva, ja que la complexitat d'aquest problema requereix un estudi aprofundit que esperem poder dur a terme més endavant.

4.1.1 Aproximacions a la segmentació del discurs

Entre les aproximacions a la segmentació dels textos, distingim entre aproximacions jeràrquiques i linears: aquestes darreres consideren els textos com una seqüència linear de segments adjacents de text, mentre que les aproximacions jeràrquiques consideren que els textos estan formats per segments que al seu torn poden contenir altres segments, de manera recursiva. Les aproximacions linears solen tractar de solucionar la segmentació de text, mentre les aproximacions jeràrquiques solen presuposar que la segmentació és una qüestió resolta i se centren en les relacions que s'estableixen entre els segments.

Aproximacions Linears

Kozima (1993) i Hearst (1994) enceten una línia de segmentació dels textos mitjançant mètodes estadístics: tracten de delimitar segments de text que tractin un mateix tema basant-se en la **similaritat lèxica** entre paraules. Des d'aquesta perspectiva, un segment queda definit com un fragment de text diferenciat de la resta perquè les paraules que el conformen són més similars entre sí que amb la resta de paraules del text. Aquesta similaritat sol calcular-se com una distància entre els diversos ítems lèxics segons la seva situació en un tesaure o un recurs lèxic similar. Els segments que s'obtenen solen tenir diversos paràgrafs d'extensió. Aquesta tècnica ha tingut un gran èxit en el processament del llenguatge natural, com mostren el gran nombre de treballs que l'utilitzen i les diverses aplicacions que ha tingut[1]. La principal objecció que es pot fer a aquesta aproximació és que tan sols reflecteix la macroestructura textual (van Dijk 1979), i s'ignoren les relacions entre oracions, frases o sintagmes.

Diversos treballs posteriors han desenvolupat aquesta proposta. Per exemple, Heinoen (1998) assenyala un dèficit en la detecció dels límits entre els segments, i presenta un mètode que incorpora

[1]Destaquem Barzilay i Elhadad (1997), Nakao (2000) i Boguraev (2000) com a principals aplicacions d'aquesta tècnica al resum automàtic.

el paràmetre de **longitud** del segment a la mesura de similaritat lèxica. Klavans et al. (1998) incorporen l'**anàlisi d'entitats** del discurs en una versió simplificada del mètode de Hearst, ja que no recorren a un tesaure per calcular la similaritat entre els termes, sinó que utilitzen la repetició de les entitats detectades, en qualsevol de les seves formes possibles.

Passonneau i Litman (1993, 1997) contrasten dos mètodes diferents de segmentació automàtica: un algorisme construït a partir de l'anàlisi d'errors i un altre basat en aprenentatge automàtic. Parteixen d'un corpus de monòlegs orals espontanis que diversos jutges humans han dividit en segments, basant-se en una vaga noció d'intenció del parlant. Analitzant aquest corpus, s'identifiquen **marques de límit de segment**, com ara relacions de co-referència entre sintagmes nominals, marcadors del discurs i marques prosòdiques. Aquestes marques se sistematitzen i s'utilitzen per crear els dos sistemes de segmentació automàtica. En els dos tipus de mètodes que contrasten, Passonneau i Litman observen que els resultats milloren de manera directament proporcional al coneixement lingüístic amb què es treballa.

Beeferman et al. (1999) també apliquen mètodes d'aprenentatge automàtic a partir d'un corpus anotat per humans. Identifiquen com a marques de límit de segment les discontinuïtats en la similaritat lèxica, però també donen compte de la contribució de certes **paraules clau**, que assenyalen de manera inequívoca un canvi de tema que les discontinuïtats lèxiques tan sols apunten. Seguint en la línia de l'aprenentatge automàtic, Hajime et al. (1998) fan una selecció dels paràmetres que millor defineixen els límits de segment, a més del de la similaritat lèxica: marcadors de topicalitat, connectors, anàfores i elisions i tipus d'oració, i determinen de manera automàtica pesos específics per a cadascun d'aquests paràmetres, amb resultats prometedors per al japonès.

Cal destacar Boguraev (2000) per la seva aportació cognitiva al concepte de segment per similaritat lèxica: el defineix com a "`allò que es troba entre punts de la narrativa que expressen una discontinuïtat perceptible`". També distingeix els tipus de segments segons si la relació que s'estableix entre ells és argumentativa o presentacional, situant-se així a mig camí entre les aproximacions linears i les jeràrquiques[2]. Aquesta proposta constitueix un salt qualitatiu a la delimitació de la noció de segment, sobretot perquè dóna paràmetres útils per la identificació humana de segments, cosa essencial a l'hora de crear un corpus anotat per jutges humans que pugui ser utilitzat en l'anàlisi o avaluació dels sistemes computacionals. Malauradament, els paràmetres que proposa resulten massa vagues per a una aplicació computacional.

Una aproximació alternativa a la delimitació dels segments és la de Wiebe i Bruce (1995), que parteixen de la informació que proporciona el **punt de vista** textual. En lloc d'utilitzar paràmetres com la similaritat lèxica o les marques prototípiques de límit de segment, identifiquen blocs en els quals totes les oracions subjectives tenen el punt de vista del mateix agent. No obstant, l'anàlisi del punt de vista requereix molt de coneixement afegit, i, com veurem tot seguit, la major part d'aproximacions que parteixen de la comprensió profunda dels textos opten per una representació jeràrquica del discurs.

Aproximacions Jeràrquiques

El concepte de segment amb què treballen les línies que acabem d'exposar es situa en un nivell de macroestructura textual (Van Dijk 1979), és a dir, que treballa amb unitats molt grans (sovint, múltiples paràgrafs). A més, aquestes unitats no presenten cap mena d'estructura entre elles, sinó que es distribueixen com a blocs successius en un text. Un nivell d'anàlisi tan general no és suficient per dur a terme una comprensió profunda del text. En àmbits com l'EI, la generació del llenguatge natural, la traducció i el resum automàtics[3], etc, cal tractar certs fenòmens que es

[2]La distinció entre aquests dos tipus de relacions es proposa en la RST (Mann i Thompson 1988), i es treballa de manera habitual en moltes aproximacions a una anàlisi estructural dels textos.

[3]Sempre que no es prengui una aproximació quantitativa.

donen en un nivell de microestructura textual: co-referència, topicalitat, relacions de coherència, etc. És per això que en els treballs que tracten l'estructura dels textos amb una gran granularitat se sol treballar amb unitats molt menors a les presentades fins ara.

En els treballs que tracten les relacions de coherència en els textos, la unitat mínima de significat textual sol ser la frase simple, tot i que no es defineix de manera clara. Mann i Thompson (1988) parteixen de la idea que les relacions retòriques s'estableixen entre 'segments de text', i sobre aquests segments tan sols diuen que no se solapen, però poden ser tan petits com un sintagma i tan grans com tot el text:

> 1) Les unitats s'han d'escollir per la seva conveniència i utilitat a la recerca en què s'inscriuen. No hi ha cap tipus d'unitat teorèticament privilegiada. (En altres paraules, per l'RST l'elecció d'unitats és pre-teorètica). S'han proposat i utilitzat oracions, sintagmes independents, paràgrafs ortogràfics i capítols de llibres.
> 2) L'elecció d'unitats és necessària per procedir amb el procés d'anàlisi.
> 3) Es considera el text com una seqüència liniar.
> 4) Les unitats escollides no s'han de solapar.
> 5) Tota elecció d'unitats ha de precedir l'anàlisi.
> A més d'això, s'espera que les unitats per cada text que s'analitza seran escollides amb algun criteri explícit aplicat de manera consistent.
>
> (Mann i Thompson 1988)

Hobbs (1985) i Hovy i Maier (1997) també se centren en establir una tipologia de relacions, però no defineixen el que és un segment. Donat que les relacions que postulen aquests investigadors són de tipus recursiu, cada segment pot contenir al seu interior més segments, de manera que s'obté una estructura jeràrquica del text.

Grosz i Sidner (1986) treballen amb segments de mida superior als que proposa la línia anterior, però no podem considerar que treballin amb poca granularitat, ja que els tracten amb una profunditat que els fa molt informatius de l'estructura textual. Utilitzen diferents paràmetres per definir diferents tipus de segments, els quals donen compte de diferents nivells d'anàlisi del text: el lingüístic, l'intencional i l'atencional. Tots els segments són susceptibles de contenir-ne d'altres del mateix tipus. En el nivell d'anàlisi que anomenen 'lingüístic', es considera com a segment 'un conjunt de frases natural'. Prenent el segment lingüístic com a base, cada segment intencional es defineix per la seva aportació diferenciada a la intenció general del discurs. En el nivell atencional, mitjançant el concepte de focus, es delimiten fragments de text que tracten els mateixos referents. Aplicat a la segmentació de textos, s'ha parat especial atenció al nivell d'anàlisi atencional, per part de les mateixes autores i també per Hahn i Strube (1997). En aquest treball, els autors parteixen del model de centratge de Grosz et al. (1995), i desenvolupen un algorisme que construeix una jerarquia de segments de discurs basats en la unitat referencial.

Yaari (1997) presenta un aprofundiment de les tècniques basades en similaritat lèxica força interessant. Malgrat que no aporta major granularitat al concepte de segment amb què es treballa, proposa una estructuració de les unitats obtingudes a partir de la tècnica de Hearst (1994). Així, no només segmenta el text d'acord amb un criteri de similaritat lèxica, sinó que forma diversos nivells d'agrupament de segments segons els diversos graus de similaritat entre ells. D'aquesta manera, s'obté una estructura jeràrquica del text. Els segments situats en nivells baixos presenten una gran similaritat lèxica interna i discontinuïtats lèxiques menors entre ells. En canvi, els segments de text que engloben aquests segments menors presenten una menor similaritat lèxica dins el bloc mateix però també una major diferència amb la resta de blocs. La representació final és indicativa dels blocs temàtics del text i la seva organització interna.

4.1.2 La nostra aproximació a la segmentació

Com hem vist, la major part de treballs sobre segmentació discursiva no treballen amb un concepte de segment que resulti útil per als nostres objectius. En alguns casos, es treballa amb unitats de baixa granularitat definides mitjançant característiques procedurals, com la similaritat lèxica, marques lingüístiques derivades de l'aprenentatge automàtic, etc., o apel·len a vagues nocions psicolingüístiques, com la 'discontinuïtat perceptible' o els 'conjunts naturals'. Pel que fa a les línies que treballen amb unitats de major granularitat, trobem a faltar una delimitació adequada de la unitat de treball, el segment. Sovint, aquesta qüestió queda rellegada a un segon pla (Mann i Thompson 1988) o es presenta com a insoluble (Marcu 1997).

Malgrat tot, creiem que, en tant que unitat de treball, cal definir el segment de manera rigorosa i adequada. A més, sovint s'utilitza aquest concepte per definir i delimitar altres elements discursius: per exemple, per a nosaltres constitueix una peça imprescindible per a distingir entre en els marcadors en funció discursiva i els que tenen funció oracional (vegeu apartat 3.2.2). És cert que una definició de segment com a unitat de discurs presenta una gran complexitat, ja que el concepte de segment està estretament lligat a qüestions encara molt difuses. La bona formació discursiva o la competència discursiva dels parlants són aspectes del llenguatge en què encara no s'ha arribat a nivells d'anàlisi satisfactòria[4]. Ens manca un aparell explicatiu que doni compte del comportament dels elements lingüístics d'aquest nivell, de la mateixa manera que ho fan, per exemple, les regles fonològiques, morfològiques, o fins i tot sintàctiques o semàntiques. Ens trobem, doncs, davant d'una complexitat que ens porta més enllà de l'abast del treball que presentem aquí. No obstant, sembla que un dels objectius del nostre treball futur ha de ser precisament una definició clara del que és segment.

També cal dir que la delimitació de segment presenta problemes pel que fa al consens entre investigadors. Sembla evident que cada treball tindrà un particular punt de vista sobre el que és segment, que estarà determinat pels seus objectius. No obstant, aquest és un problema freqüent en lingüística: trobem que tampoc no hi ha una definició globalment acceptada d'unitats com la paraula o l'oració, que compten amb una tradició d'estudi important.

Presentem tot seguit una primera aproximació a la definició del concepte de segment. Malgrat que no hem assolit una definició satisfactòria, hem tractat de delimitar en certa mesura aquest element per tal de poder desenvolupar el nostre treball aplicat: l'estudi sobre marcadors del discurs i el segmentador retòric. Alhora, aquest treball aplicat ha constituït també la comprovació empírica del concepte de què partíem, contribuïnt a refinar-lo i adequar-lo a la realitat dels textos. El treball de Polanyi (1988, 1995), ha constituït un punt de referència bàsic en aquesta tasca, ja que supera la major part de les mancances de les aproximacions que hem esmentat fins ara. En el model del discurs de Polanyi (1995) se segmenta el text en unitats lingüísticament rellevants des del punt de vista sintàctic, proposicional i discursiu, de manera que un segment de text queda definit així:

```
Un Constituent del Discurs Elemental (Elementary Discourse Constituent Unit, DCU)
és una expressió mínima que transmet un únic event o estat indexat pel que fa a la
situació física i social de l'expressió (ja sigui real o modelitzada), a la modalitat,
la polaritat, l'especificitat, el punt de vista, etc.  Un DCU pot ser una frase o
oració simple però també pot no ser-ho.  Es tradueixen com a expressions quantificacionals
bàsiques.
```

Amb un conjunt de regles recursives, aquestes unitats es combinen formant altres unitats que les contenen i que, al seu torn, poden estar contingudes en altres unitats, donant lloc a una estructura jeràrquica del discurs. Per dur a terme aquesta tasca cal una profunda anàlisi sintàctica, semàntica

[4] És per això que, sovint, emprem un terme tan vague com 'felicitat' o 'infelicitat' per referir-nos a la noció de bona formació discursiva.

i discursiva, cal recórrer a fonts de coneixement lingüístic molt extenses i computacionalment molt costoses. És per això que, malgrat la seva vàlua com a model teòric, aquesta proposta no és implementable en la major part de sistemes actuals.

Basant-nos en el treball de Polanyi, proposem una definició de segment com a base operativa per recolzar el nostre treball:

> un **segment simple** `és igual o major a una proposició, és a dir, del qual es pot predicar un valor de veritat. És una unitat mínima de significat discursiu, és a dir, no conté cap altre segment o relació, i alhora és susceptible d'establir una relació de tipus textual (de cohesió, de coherència, etc.) amb un altre segment.`
>
> un **segment compost** `és un segment que en conté dos o més relacionats entre ells per una o més relacions de tipus textual. Els segments continguts dins d'un altre segment poden establir una relació simètrica o asimètrica entre ells.`

El fet que els segments estan relacionats entre ells és el que fa possible l'existència de segments compostos, i, alhora, fa possible que puguem obtenir una estructura no linear del text. Podem donar compte d'aquesta estructura mitjançant regles recursives que estableixin el funcionament de les relacions entre segments. Aquestes relacions entre segments solen estar fortament caracteritzades pel seu contingut retòric. No obstant, determinar el contingut retòric de cada relació o de cada marcador és una tasca complexa i que requereix un esforç important d'investigació. Per això creiem que, en un estadi primerenc del nostre treball, podrem donar compte de l'estructura retòrica dels textos sense contemplar el contingut de les relacions, centrant-nos en la seva vessant purament estructural. Tal com apunten Mann i Thompson (1988), la relació retòrica més característica és la de nuclearitat, ja que en la major part de relacions entre segments es dóna que un dels segments és el nucli, mentre que un altre li està retòricament subordinat. Creiem que, a banda del contingut retòric específic de cada relació entre segments, si treballem sobre la relació de nuclearitat podrem obtenir una estructura del text. Encara que aquesta estructura no reflectirà el contingut, podem treballar-hi per obtenir un resum automàtic.

Hem de tenir en compte que la funció principal de la definició de segment que hem proposat és establir una base teòrica sòlida per al nostres objectius aplicats. Ara bé, hi ha fragments de text que no s'adiuen amb la definició de segment però, no obstant, els percebem com a discursivament prescindibles, i, per tant, obviables de cara a l'elaboració d'un resum d'un text. Així doncs, la principal mancança de la definició que hem proposat és que no dóna compte de certs segments útils per a resum. Si bé és cert que aquests elements no poden funcionar de manera independent en un text, també és veritat que, per al resum automàtic, és útil de detectar-los i tractar-los, de la mateixa manera que es tracten els segments que sí s'adiuen amb la definició. Aquests segments són, majoritàriament, adjunts de tipus circumstancial no requerits per cap component oracional, i també fragments de text entre parèntesis[5].

exemple: ***Con gorra y camiseta alusivas a la fecha y banderas de papel en mano****, millares de panameños marcharon hasta Balboa desde distintos e históricos puntos de la ciudad.*

> En aquest exemple, el fragment en negreta no és necessari per a la bona formació discursiva de l'oració i podria ser-ne eliminat sense que això suposés un canvi de significat o agramaticalitat. No obstant, aquest fragment de text no constitueix una proposició per ell mateix.

exemple: *De hecho, algunos varones más moderados de la dirección del partido han alertado contra un espíritu de venganza apresurada y hasta sugieren que una crudeza extrema con el antiguo líder podría ser muy mal recibida por un sector* ***(tierno y viejo)*** *de la sociedad.*

[5] No considerem les aposicions en aquest grup perquè creiem que de qualsevol element substantiu se'n pot predicar l'existència, de manera que constitueix una proposició per ell mateix.

En aquest exemple, el fragment de text entre parèntesi podria ser eliminat de l'oració sense pèrdua de significat o gramaticalitat, però no podem dir que consitueixi una proposició per ell mateix. En general, gairebé tots els segments entre parèntesi són obviables, independentment que consitueixin una proposició o no.

Hi ha marques explícites en els textos que semblen reforçar la idea que aquests fragments de text constitueixen segments amb entitat discursiva pròpia: signes de puntuació, peces lèxiques i estructures sintàctiques. Esperem poder sistematitzar aquestes característiques per arribar a una delimitació satisfactòria de la nostra unitat de treball. Tot seguit presentem les conclusions provisionals a què hem arribat i a partir de les quals hem desenvolupat una eina aplicada, el segmentador retòric.

En la nostra aproximació a la segmentació retòrica hem partit del supòsit que és possible i rellevant d'identificar segments a l'interior de les oracions, ja que una de les principals limitacions dels sistemes de resum automàtic basats en l'extracció de fragments literals del text és que prenen oracions completes, quan sovint tan sols una part de l'oració és realment rellevant. En aquest context, sembla evident que algunes frases subordinades han de ser considerades com a segments, ja que són proposicions de ple dret. Efectivament, si apliquem la prova que proposa Knott per identificar marcadors (vegeu apartat 3.1.1.) a la inversa, podrem observar com, en eliminar l'element que considerem marcador d'una frase subordinada, la frase esdevé una oració independent, ja que, habitualment, allò que marca la seva relació discursiva amb un altre segment marca també la seva subordinació sintàctica.

Però anem més lluny: considerem, com Schauer i Hahn (2000), que alguns sintagmes preposicionals, no dominats per la valència verbal de la oració en què es troben, poden fer una aportació retòrica al discurs. És cert que, des d'un punt de vista tipològic, en llengües amb una gran tendència a la nominalització, aquest fenomen sembla més justificable, ja que els sintagmes preposicionals que considerem eliminables contenen un element nominal d'origen verbal, i podem considerar que aquest element nominal constitueix el predicat d'una proposició. Si extrapolem aquest fenomen a la resta de construccions sintàctiques, observem que, efectivament, un constituent oracional sembla tenir una entitat discursiva suficient per ser considerat com a segment si s'adiu amb la definició que hem proposat però, a més, no està subcategoritzat per un predicat. Així, considerarem com a segments les frases subordinades, els sintagmes preposicionals, els sintagmes adjectius, els sintagmes adverbials... no dominats per la valència de cap constituent oracional, tot i que, en l'estat actual de la nostra investigació, tan sols tractem els que fan una aportació discursiva clara. Considerem que fan una aportació discursiva clara els components oracionals amb un contingut retòric explícit que contemplem en una tipologia provisional de relacions retòriques: de causa, de contraargument, de condició, etc., ja siguin frases subordinades, sintagmes preposicionals, sintagmes adjectius o grups adverbials que sistemàticament aporten algun d'aquests significats.

exemple: *[En el trabajo que le mereció el Nobel],* ***[Laughlin] demostró****, [junto con Horst Störmer y Daniel Tsui],* ***[que en determinadas condiciones los electrones pueden formar nuevos tipos de partículas cuya carga eléctrica es exactamente un tercio de la carga del electrón]****.*

En aquest exemple, les parts amb negreta són les que estan subcategoritzades pel predicat 'demostró', mentre que els segments entre claudàtors que no estan en negreta, no estan subcategoritzats.

Les dues primeres fonts d'evidència s'han utilitzat de manera sistemàtica en els treballs que hem mencionat en l'apartat anterior. Ara bé, s'ha treballat poc la interacció entre estructura retòrica i estructura sintàctica, malgrat que sembla clar que existeix una relació entre la segmentació discursiva d'alta granularitat i l'anàlisi sintàctica de nivell oracional (Polanyi 1995).

Així doncs, partint de l'oració com a segment bàsic, hem considerat que les cadenes de paraules d'una oració no caracteritzades per cap marcador amb funció discursiva constitueixen el cos central de l'oració. Els marcadors del discurs determinen el tipus de relació que s'estableix entre aquest cos central i els segments caracteritzats per algun marcador. Hem distingit dos tipus de relació entre els segments d'una oració:

dependència un segment és dependent del cos central d'una oració quan no pot constituir una oració autònoma sense transformacions sintàctiques considerables. Solen establir relacions de dependència els segments constituïts per sintagmes preposicionals, sintagmes adjectius, sintagmes adverbials, aposicions, frases de relatiu i frases impersonals. Perquè aquests segments poguessin funcionar com una oració sintàcticament completa caldria afegir-los un o més constituents oracionals.

> ***exemple:*** *El vicepresidente del Gobierno,* ***[Francisco Alvarez-Cascos]****,* ***[recluido toda la noche en el búnquer de la Moncloa]****, aceptó que quizá se había exagerado, pero arguyó que era preferible haberse "pasado,* ***[entre comillas]****, por el lado de la seguridad que por el otro".*
>
> En aquest exemple, considerem els segments en negreta com a dependents de la resta de la oració.

independència un segment és independent si pot constituir una oració autònoma mitjançant la prova que hem presentat més amunt, és a dir, si tenen un predicat amb totes les restriccions de selecció satisfetes. En general, això es dóna en els segments caracteritzats per un marcador del tipus 'subordinant' o 'conjunció'.

> ***exemple:*** *El vicepresidente del Gobierno, Francisco Alvarez-Cascos, recluido toda la noche en el búnquer de la Moncloa, aceptó que quizá se había exagerado,* ***pero arguyó que era preferible haberse "pasado, entre comillas, por el lado de la seguridad que por el otro"****.*
>
> En aquest exemple, considerem qe el segment en negreta és independent de la resta de la oració, amb qui estableix una relació retòrica d'adversació.

D'aquesta manera, hem arribat a una noció operativa de segment com a unitat discursiva bàsica. Basant-nos en aquest concepte hem desenvolupat una aplicació informàtica que presentem tot seguit.

4.2 Una aplicació: El Segmentador Retòric

Hem dissenyat i implementat una eina de segmentació retòrica automàtica per a l'espanyol que vol ser independent de domini. Hem sistematitzat el coneixement de què disposàvem per tal de detectar i explotar les marques superficials de segmentació en els textos. Així, hem aplicat la noció de segment que acabem de presentar i la informació obtinguda a partir de l'estudi de marcadors del discurs del capítol 3. En aplicar aquesta informació als textos reals, hem comprovat la seva adequació i hem introduït les modificacions necessàries en les definicions i en la descripció de marcadors.

L'objectiu principal d'aquesta anàlisi segmental és el pre-procés del text per a la interpretació de l'estructura retòrica dels textos, la qual podrà ser utilitzada en l'elaboració d'un resum automàtic. No obstant, l'anàlisi segmental dels textos pot aportar informació valuosa en altres camps del PLN, com ara l'anàlisi d'entitats o l'anàlisi sintàctica.

4.2.1 Anàlisi de corpus

En l'estudi sobre corpus dels marcadors del discurs (apartat 3.2.1), hem detectat aquella informació segmental que tenen associada. Així, hem pogut perfilar concepte inicial de segment amb què treballem, i hem configurat la descripció necessària dels marcadors del discurs en tant que marques de segmentació. En aquest estudi, hem trobat les següents característiques dels marcadors del discurs que aporten informació explícita sobre els segments:

- els **signes de puntuació** solen ser marques de límit de segment
  - els **punts** (i seguit, d'admiració, d'interrogació, suspensius, punt i coma, dos punts) són marques inambigües de límit de segment independent
  - les **comes** són marques ambigües de límit de segment, ja que no és clar si delimiten segments independents o dependents, a més, poden formar part d'un segment amb estructura de llista
  - els signes **parentètics** (inclosos els guionets que fan la mateixa funció) són marques inambigües de límit de segment, però no podem determinar si es tracta d'un segment dependent o independent
  - les **cometes** són marques inambigües de límit de segment, però no podem determinar si es tracta d'un segment dependent o independent
- els **marcadors del discurs lèxics**
  - combinats amb coma o punt, són marques inambigües de límit de segment independent, és a dir, tenen una funció discursiva inambigua
  - sense coma o punt, poden ser
    * si són marcadors de funció discursiva, són marques inambigües de límit de segment independent
    * si són marcadors de doble funció (poden tenir una funció discursiva o oracional), són marques ambigües de límit de segment, ja que poden delimitar un segment independent o poden introduïr constituents subcategoritzats.
- algunes **estructures sintàctiques**
  - les frases de **relatiu**, precedides de coma, són marques inambigües de segment dependent
  - les frases de **gerundi** o **participi**, precedides de coma o punt, són marques inambigües de segment independent.

La definició de segment que hem proposat en els apartats 4.1.2 i 4.1.3 recorre a conceptes com 'proposició' o 'subcategorització' que tan sols poden ser utilitzats computacionalment mitjançant una anàlisi profunda dels textos. Donat que no disposem dels recursos de PLN necessaris per assolir la profunditat d'anàlisi requerida per treballar amb aquests conceptes, proposem una segona versió d'aquesta definició, basada en el comportament de les marques superficials de segmentació que hem trobat en l'estudi de corpus:

> `Considerem` **segment** `tota cadena de paraules entre dues marques inambigües de límit de segment.`

Malgrat que aquesta definició està fortament determinada per les observacions de corpus i per les característiques de l'anàlisi automàtica que volem realitzar, creiem que és lingüísticament adequada perquè intenta recollir l'essència del concepte de segment tal com el formulàvem des de la perspectiva lingüística, però adaptant-lo a les limitacions dels recursos de PLN de què disposem.

Hem constituit un corpus de treball per a desenvolupar una primera versió de la nostra eina. Aquest corpus pertany al gènere periodístic, ja que considerem que els textos literaris, col·loquials, d'especialitat, etc. contenen un gran nombre d'irregularitats o de construccions molt particulars que poden tractar-se de manera més adequada un cop s'ha establert un estàndard. En aquest sentit, el gènere periodístic ens ofereix corpus abundants i propers a un registre estàndard, no marcat, de la llengua, per la qual cosa és molt adequat per a l'estudi de les regularitats discursives, sense oblidar mai que, com a gènere que és, s'hi donen també certes regles particulars. Esperem que, en versions posteriors del segmentador, puguem incorporar-hi els resultats de l'estudi d'altres gèneres, per tal de garantir una major transportabilitat de l'eina. També pot resultar interessant desenvolupar versions més adequades a gèneres o registres particulars.

El corpus de treball, amb un volum aproximat de 100.000 paraules, està format per fragments aleatoris del corpus de treball que hem utilitzat per a l'estudi de marcadors del discurs. Aquest ha estat el banc de proves del segmentador, en què hem anat provant les diverses modificacions i millores que incorporàvem a l'eina. Hem establert un segon corpus periodístic, de 20.000 paraules, independent del primer banc de proves, sobre el qual hem realitzat el refinament de l'eina. També hem establert un altre petit corpus independent del corpus de treball inicial, de 3.000 paraules, per a l'avaluació humana dels resultats de segmentació.

4.2.2 Disseny, Implementació i Resultats

El segmentador ha estat dissenyat per reconèixer automàticament els segments tal com els hem establert en l'apartat anterior. Disposa d'informació morfosintàctica, perquè treballa sobre un text analitzat mitjançant l'analitzador morfològic MACO (Arévalo et al. 2001), el desambiguador d'etiquetes morfològiques RELAX (Padró 1997) i l'anàlisi sintàctica parcial de la gramàtica d'Atserias et al. (1998a). Aquesta informació es compatibilitza amb la informació pròpiament segmental sorgida de l'anàlisi de corpus que hem exposat en l'apartat anterior. L'eina està implementada en forma de gramàtica independent de context en l'analitzador TACAT (Atserias et al. 1998a), seguint el model de gramàtica de Civit i Castellón (1998).

Hem optat per utilitzar tan sols la informació segmental inambigua, de manera que es prioritza la precisió per sobre de la cobertura, és a dir, que es reconeixen menys segments que no reconeixeria un humà però amb la garantia que la major part de segments que es reconeixen són efectivament segments.

Les unitats detectades pel segmentador són:

- **paràgraf** - tota cadena d'oracions entre dos salts de línia.
- **oració** - tota cadena de delimitadors i segments entre dos punts.
- **delimitadors** - marques de límit o abast dels segments:
  - **pauses sintàctiques**
    * inambigües: els punts sempre marquen el final d'un segment i l'inici d'un altre, els signes parentètics i les cometes sempre marquen l'inici i final d'un element parentètic, que pot ser segment o no.
    * ambigües: les comes. Són inambigües associades a marcadors del discurs: marquen el límit inicial i/o final de segments caracteritzats per algun marcador del discurs.
  - **marcadors del discurs:** En l'estat de desenvolupament actual de l'eina, tan sols tenim en compte com a delimitadors els marcadors sintàcticament incomplets (subordinants i preposicions), ja que no hem pogut sistematitzar la distribució dels marcadors sintàcticament complets (adverbials i anafòrics), i per tant no podem determinar de manera inambigua la seva funció delimitadora.

* inambigus: els marcadors que no poden estar dominats per la valència d'un constituent oracional.
* ambigus: els marcadors que poden estar dominats per la valència d'un constituent oracional. Són inambigus associats amb certs patrons de pausa sintàctica: algunes conjuncions subordinants, les conjuncions coordinants i algunes locucions adverbials, precedides de coma o punt, marquen el final d'un segment i l'inici d'un altre.

– **estructures sintàctiques**

* inambigües: constituents oracionals introduïts per preposicions que no poden estar dominades per la valència d'un constituent oracional.
* ambigües: associades amb certs patrons de pausa sintàctica: les frases de relatiu, frases no personals, sintagmes preposicionals, sintagmes adverbials, i alguns tipus de sintagmes adjectius i d'aposició, sempre que es donin entre pauses sintàctiques, constitueixen alhora que delimiten segments.

– **segments** - identifiquem com a segments tota cadena de paraules entre delimitadors, acomplint els patrons establerts en la gramàtica per cada tipus de delimitador. Distingim:

– segments **dependents** - frases de relatiu, sintagmes preposicionals, sintagmes adverbials i els casos que reconeixem de sintagmes adjectius i d'aposició, sempre entre pauses sintàctiques. Distingim els segments i els considerem com a dependents del segment en què es troben.

– segments **independents** - tota cadena de paraules entre dos punts o bé tot segment caracteritzat per un marcador del discurs subordinant o no personal inambigu o en context inambigu.

– segments **possibles** - tot fragment de text caracteritzat per un delimitador del discurs ambigu, és a dir, amb una possible funció oracional, excepte si es tracta d'estructures sintàctiques.

Tot seguit descriurem breument el **funcionament** del segmentador. En l'Annex I adjuntem una descripció completa de totes les regles sintàctiques del segmentador, amb alguns exemples il·lustratius del funcionament d'algunes regles i dels fenòmens de què donen compte. En l'Annex II presentem una mostra de text segmentat, en què presentem i discutim un dels problemes que presenta: la solució dels segments de tipus 'coordinat'.

Per raons de processament, les unitats de treball del segmentador es caracteritzen per paràmetres diferents dels que hem llistat més amunt, que ens interessen per a l'anàlisi retòrica dels textos. Per començar, distingim diferents tipus de segment, segons la seva composició:

– **segments amb format**, aquells que estan caracteritzats per algun tipus de marcador del discurs, ja sigui marcador lèxic o estructura sintàctica.

– **segments forts**, aquells que estan caracteritzats per un marcador del discurs amb funció delimitadora discursiva inambigua

– **segments febles**, aquells que estan caracteritzats per un marcador del discurs amb doble funció, discursiva o oracional

– **segments sense format**, les cadenes de paraules no caracteritzades per cap marcador del discurs.

– **segments parentètics**, aquells que estan delimitats per signes parentètics o cometes. No els incloem entre els segments amb format perquè la seva distribució no és sistemàtica com la dels segments amb format, sinó que és aleatoria com la de les paraules sense funció discursiva determinada.

També distingim tipus diferents de delimitadors, segons el context de pauses sintàctiques que requereixen per la seva caracterització inambigua de marcadors amb funció de delimitació discursiva:

- **requereixen pausa sintàctica** els marcadors del discurs amb doble funció, ja siguin peces lèxiques o estructures sintàctiques (entre comes i/o punts i/o delimitadors inambigus):
  - la major part de sintagmes preposicionals i frases subordinades, a no ser que estiguin introduïts per un marcador amb funció discursiva inambigua
  - la major part de sintagmes adverbials, a no ser que es tracti de marcadors del discurs o locucions adverbials amb funció discursiva inambigua
  - els sintagmes adjectius
  - les frases no personals
  - les frases de relatiu
  - les frases coordinades, ja que les conjuncions coordinants poden unir unitats inferiors a un segment (per exemple, membres de llistes)(vegeu Annex II)
- **no requereixen pausa sintàctica** els marcadors del discurs amb funció discursiva inambigua.

A partir d'aquestes distincions hem trobat patrons sistemàtics de distribució dels segments que ens permeten determinar amb seguretat quan un marcador caracteritza de manera inambigua un segment de tipus fort. A partir de la unitat 'oració', distingim tres posicions en què es donen comportaments segmentals sistemàtics i diferenciats, sobretot pel que fa a la distribució dels segments amb format:

mig: en la posició central d'oració donem compte de la recursivitat segmental de les oracions, alhora que establim el mínim necessari perquè es constitueixi una oració: una cadena de com a mínim una paraula que pot tenir un segment de tipus parentètic en qualsevol posició, de manera que donem compte de la distribució aleatòria dels segments parentètics. Les regles de segmentació en aquesta posició donen com a enniuats els segments amb format, de manera que dues cadenes de paraules sense format a banda i banda d'un segment amb format delimitat per comes constitueixen un únic segment.

exemple: ***Si es seguro que la recuperación del SPD tiene que ver con los gestos del canciller Schröder****, [que, [sabiamente], ha dejado de ir de coteórico de Tony Blair en la Tercera Vía y está a pie de obra en Berlín],* ***no lo es menos que el affaire Kohl ayuda y mucho****.*

En aquest exemple, la frase de relatiu entre claudàtors és un segment enniuat en el segment 'mig', que, en aquest cas, abasta la totalitat de l'oració.

Donem compte de la recursivitat segmental com una seqüència de comes, segments amb format i segments sense format, de manera que els segments amb format de tipus fort tan sols poden donar-se en un context inambigu, és a dir, sempre delimitats per comes o delimitadors inambigus. Això és així perquè, com ja hem dit, en el cas dels marcadors amb doble funció tan sols podem establir de manera inambigua la seva funció com a delimitadors inambigus si estan precedits de coma. És per això que, si un segment amb format està caracteritzat per un marcador amb doble funció, mai no pot donar-se immediatament a la dreta de la coma que tanca un segment amb format, perquè en aquesta posició pot estar subcategoritzat per un component oracional fora del segment immediatament anterior.

exemple: *"El dinero invertido ha servido,* ***[además de para que no ocurriera nada]****, para disponer de equipos informáticos más modernos, más rápidos y servicios mejores"*

En aquest exemple, si considerem que darrera d'un segment amb format podem trobar un altre segment amb format, estarem categoritzant com a segment una part de l'oració que està subcategoritzada pel predicat.

Ara bé: en l'estat actual de desenvolupament del segmentador, no disposem de dades que ens permetin determinar en quins casos un segment amb format introduït per un marcador amb doble funció que es dóna immediatament a la dreta d'un altre segment amb format serà subcategoritzat o no, i és aquesta la principal font d'error del segmentador.

exemple: *Los Quince deberían alcanzar un acuerdo político antes de que finalice el año, según lo pactado en la cumbre de Feira, en la que se logró que Luxemburgo aceptara, tras un periodo transitorio de siete años, el intercambio de información sobre los ahorradores, lo que se traduce en acabar con el denominado secreto bancario.*

L'anàlisi humà d'aquesta oració seria:

Los Quince deberían alcanzar un acuerdo político antes de que finalice el año, [según lo pactado en la cumbre de Feira, [en la que se logró que Luxemburgo aceptara, [tras un periodo transitorio de siete años], el intercambio de información sobre los ahorradores, [lo que se traduce en acabar con el denominado secreto bancario]]].

Però l'anàlisi automàtica és:

Los Quince deberían alcanzar un acuerdo político antes de que finalice el año, *[según lo pactado en la cumbre de Feira],* ***en la que se logró que Luxemburgo aceptara,*** *[tras un periodo transitorio de siete años],* ***el intercambio de información sobre los ahorradores,*** *[lo que se traduce en acabar con el denominado secreto bancario].*

Sembla que la manera d'arribar a donar compte d'aquest tipus de fenòmens passa per l'anàlisi de dependències entre constituents oracionals, però això requereix una anàlisi profunda que queda fora del nostre abast. Una altra possible via per solucionar, en certa mesura, els errors d'aquest tipus seria la sistematització de patrons de seqüències de segments de manera que el context contribueixi a determinar les dependències entre segments, i ampliar els factors contextuals que ens ajuden a determinar de manera inambigua quan un marcador té funció de delimitador i quan no la té.

Ja que els segments amb format successius presenten problemes de cara a la seva caracterització inambigua, hem pres l'opció que semblava més freqüent en els textos, i hem optat per considerar que tot segment amb format immediatament a la dreta d'un segment enniuat no serà un segment enniuat a no ser que es tracti dels segments que hem categoritzat com a 'dins', segments que poden trobar-se enniuats dins d'un segment amb format: sintagmes adverbials entre comes, sintagmes adjectius entre comes, noms propis entre comes o sintagmes preposicionals entre comes, tot i que aquest darrer grup és el que presenta més problemes.

inicial: en inici d'oració[6] es dóna un context de pausa sintàctica que fa inambigus els delimitadors ambigus. No obstant, no es poden donar segments amb format de tipus adjectiu ('relatiu', 'aposició' o 'adjectiu'), ja que el format que els caracteritza és ambigu en aquesta posició: el mateix format es dóna per components nominals subcategoritzats.

exemple: ***Agotado el sistema,*** *el beneficiario ha sido un ex golpista, el presidente Hugo Chávez, patrocinador de un populismo que mira a Cuba.*

En aquest exemple, la frase de participi és inambiguament no subcategoritzada

[6] Es considera inici d'oració: principi de text, principi de paràgraf, text immediatament darrera de punt.

En aquesta posició també podem considerar que es poden encadenar de manera inambigua dos segments amb format, a diferència del que passa en la posició central d'oració. Això és així perquè la subcategorització tendeix a projectar-se cap a la dreta, i el segments en posició inicial d'oració es troben a l'esquerra dels components oracionals que podrien subcategoritzar-los i, per tant, alterar el seu estatus com a segment. En canvi, sí que és possible que aquests segons segments amb format en la posició inicial d'oració estiguin subcategoritzats pels segments de la seva dreta, per la qual cosa resulta especialment adequat reconèixer-los com a segments amb format i tractar-los com a tals. La única excepció, tot i que molt rara, d'aquesta norma, la donen els segments amb format del tipus no personal separades per comes de la part central de l'oració[7]:

exemple*: *Agotado, *es lo que estoy.*

> En aquest exemple, la dislocació a l'esquerra d'un element subcategoritzat es marca amb una pausa sintàctica, que és alhora la marca formal a partir de la qual el segmentador troba el límit final dels segments amb format. Donat que aquest element dislocat presenta el mateix patró que un segment de tipus no personal (un participi limitat per una coma), l'anàlisi que el segmentador dóna d'aquesta oració l'etiqueta com a un segment diferenciat de la resta de l'oració, i no com a part integrant.

Segons les regles que acabem d'enunciar, aquestes seqüències quedarien caracteritzades incorrectament. Esperem poder aprofundir el nostre estudi de corpus sobre aquests casos per poder detectar i sistematitzar marques formals que ens permetin de corregir aquesta mancança.

final: com en principi d'oració, en final d'oració, és a dir, immediatament abans del punt que tanca un segment del tipus 'oració', es dóna un context de pausa sintàctica que fa inambigus els delimitadors ambigus, i es dóna el context suficient i necessari perquè ocorrin certs tipus de segment que tan sols ocorren en aquesta posició: segments parentètics delimitats per un guionet i un punt. No obstant, en aquesta posició no es dóna el mateix fenòmen característic de la posició inicial d'oració: no podem trobar dos segments amb format concatenats dels quals el segon estigui inequívocament lliure de la subcategorització verbal, si bé és cert que hem detectat tendències generalitzades que semblen apuntar que en aquesta posició la subcategorització és més feble que en la posició central d'oració. Esperem poder aprofundir el nostre estudi de corpus per poder assegurar aquestes tendències amb més fermesa.

4.2.3 Avaluació de l'eina

Hem intentat avaluar el funcionament del segmentador en dues vessants: contrastant els resultats amb els objectius inicials per una banda, i amb un corpus anotat per jutges humans imparcials, per un altra. En primer lloc, però, hem donat compte de les limitacions de què partim per assolir aquests objectius inicials.

Limitacions

Per avaluar en la seva justa mesura el funcionament del segmentador, cal tenir en compte que treballa sobre els resultats d'una sèrie de processos de PLN realitzats mitjançant altres eines. Per tant, el segmentador s'ha d'adequar a les característiques d'aquestes eines i assumir-ne les mancances. Les característiques que les afecten de manera més directa el funcionament del segmentador són:

– en la identificació i tractament de les **entitats** del discurs (majoritàriament, noms propis), s'observen mancances en:

[7] Això és així per limitacions de la desambiguació d'etiquetes morfològiques en què ens basem: les conjuncions subordinants i els relatius amb forma 'que' reben sovint la mateixa etiqueta morfològica, i tampoc no es pot distingir sempre entre els participis amb funció adjectiva o amb funció verbal.

	percentatge	nombre total de segments (292)
COBERTURA	**73'63%**	215
SILENCI	**26'73%**	77
PRECISIÓ	**95'89%**	203
SOROLL	**4'11%**	12

Taula 4.1: Resultats Quantitatius del Segmentador

  - identificació insuficient d'índexs temporals: es reconeixen dates però no es reconeixen índexs com *la próxima semana*, *el año pasado*.
  - segmentació deficient d'alguns noms propis

- pel que fa a la desambiguació **morfològica**, trobem que:

  - darrera de coma, les conjuncions subordinants completives reben etiqueta de pronoms relatius, per la qual cosa no podem distingir adequadament components oracionals subcategoritzats i frases de relatiu.

- referent a l'anàlisi **sintàctica**, l'eina que emprem realitza una anàlisi parcial, cosa que suposa uns límits que al seu torn limiten també l'anàlisi retòrica posterior:

  - no es reconeixen les coordinacions o juxtaposicions a nivell sintagmàtic
  - no hi ha informació sobre dependències sintàctiques a nivell oracional ni a nivell de constituents sintagmàtics (tret d'adjectius)

Tenint en compte aquestes contingències derivades de l'estat actual de les eines de processament del llenguatge que emprem, l'avaluació del segmentador mostra com a principals limitacions pròpies:

- una gran **dependència dels signes de puntuació**, cosa que implica
- **identificació insuficient de delimitadors** amb funció de delimitació discursiva inambigua, ja que sovint, si no co-ocorren amb patrons de puntuació molt definits, no es pot determinar la funció d'un marcador del discurs amb doble funció, cosa que implica
- **identificació insuficent de segments** de tipus fort, ja que si no s'identifica un delimitador de manera inambigua, no es pot identificar el segment que delimita

Així doncs, reconeixem menys segments dels que realment hi ha en un text. Això és així perquè prenem una postura conservadora vers la segmentació, és a dir, que intentem identificar els segments quan tenim prou marques formals que en garanteixen la presència. També intentem garantir que la informació de tipus de segment serà gairebé sempre correcta. Quan podem determinar el tipus de segment, podem també determinar si estableix una relació de dependència o independència amb el nucli oracional.

Tenint en compte, doncs, que ens trobem encara en fase de desenvolupament, presentem una breu avaluació dels resultats d'aquesta versió inicial del segmentador.

Resultats

Hem etiquetat manualment un corpus de 3000 paraules intentant reflectir els resultats òptims d'una anàlisi amb el segmentador automàtic. Aquest corpus d'avaluació és de gènere periodístic i domini divers, com els corpus que hem utilitzat en les proves del segmentador, però no està inclòs en aquest primer corpus de treball. El segmentador ha identificat correctament un 76'63% dels segments que havíem etiquetat prèviament en el corpus, no ha identificat un 26'73% dels segments i ha identificat de manera errònia un 4'11% dels segments (vegeu taula 4.1). Així doncs, el segmentador cobreix tres quartes parts dels segments d'un corpus (mesura de cobertura), i, per tant, no dóna compte d'una quarta part (mesura de silenci). L'índex d'error en la identificació de segments (mesura de soroll) és significativament baix, per la qual cosa la mesura de precisió és molt alta.

El baix índex de soroll que hem registrat i un valor relativament alt pel que fa al silenci reflecteix clarament la nostra postura conservadora respecte del reconeixement de segments. D'altra banda, hem de dir que aquests resultats estan en procés de millora, ja que hem identificat les principals tendències d'error i esperem poder corregir-les breument. La major part d'errors venen donats per:

- gestió insuficient de qüestions de **format** de document, com ara el títol, l'autor, l'agència de notícies, la localització de la notícia, la data, etc., i també la presència de símbols desconeguts per al segmentador (|, », [, etc.)[8]
- cobertura insuficient de **marcadors** del discurs
  - alguns marcadors **no estan recollits** en la base de dades que utilitza el segmentador
  - encara no podem donar compte de la funció delimitadora d'alguns marcadors de **tipus sintàctic complet**, tot i que esperem incorporar aquesta informació més endavant.
- hem de donar compte d'algunes **interaccions entre regles** que donen errors:

  exemple: *oracio_[coordinat_[Pero también hay casos en la física] , según Laughlin , como la superconductividad a alta temperatura , rel_[que sigue inexplicado] :]*

  Sembla que, aplicant les regles de la gramàtica de segmentació 65, 28 i 32 (vegeu Annex I), l'anàlisi hauria de ser:
  oracio[coordinat_[Pero también hay casos en la física, prepo_[según Laughlin], como la superconductividad a alta temperatura, rel_[que sigue inexplicado]:]

- cal delimitar més adequadament l'abast dels segments del tipus **'coordinat'**, que són marcadament diferents de la resta de segments amb format (vegeu Annex II).

  exemple: *oracio_[Es una obra de la que , feble_[si se acepta su juego] , más vale huir ;]oracio_[coordinat_[pero que] , feble_[si se acepta] , conduce a rincones inéditos del lenguaje cinematográfico]*

  L'anàlisi desitjable d'aquesta oració seria:
  oracio_[Es una obra de la que, feble_[si no se acepta su juego], más vale huir];
  oracio[coordinat_[pero que, feble_[si se acepta], conduce a rincones inéditos del lenguaje cinematográfico]

- identificació inadequada de **cadenes de segments amb format**, tot i que creiem que podrem donar compte de certes seqüències típiques de segments amb format que poden aportar el context suficient per desambiguar marcadors del discurs fins ara ambigus, com per exemple frases subordinades de relatiu, grups preposicionals, etc.
- identificació insuficient de segments de tipus **'aposició'**, **'prepo'**, **'adjec'** i **'adver'**, per limitacions de les eines de PLN que processen el text d'entrada del segmentador. Si donem compte de certs patrons segmentals, com hem dit en el punt anterior, podrem aportar cert context que pot contribuir a determinar de manera més acurada aquests tipus de segment amb format sintàctic.

[8] En articles periodístics, aquestes qüestions solen ser la causa d'entre el 5% i el 20% dels errors de segmentació.

Així doncs, els resultats del segmentador aconsegueixen de donar compte d'un bon nombre de segments intraoracionals. El seu índex d'error resulta acceptable, i esperem que millorarà quan corregim les tendències d'error que hem detectat. D'altra banda, l'índex de cobertura encara ha de ser millorat, la qual cosa requereix un esforç d'investigació major. No obstant, creiem que millorarem la cobertura d'aquesta eina a mesura que aprofundim en les qüestions teòriques de segmentació que queden per resoldre. Així mateix, la incorporació d'altres tipus d'informació pot resultar molt útil per a la solució de moltes qüestions de segmentació que ara queden fora del nostre abast. La detecció i seguiment de les entitats d'un text, la col·laboració entre informació sintàctica i segmental, la interacció entre segmentació i interpretació retòriques... poden ser mútuament enriquidores, millorant alhora la segmentació però també els altres processos del llenguatge.

Avaluació humana

Per tal de garantir l'adequació dels resultats de l'eina a la realitat lingüística, és metodològicament recomanable recórrer a jutges humans aliens al desenvolupament de l'aplicació. Aquests jutges han de reflectir una referència no esbiaixada. És per això que hem dissenyat un experiment per recollir les intuïcions dels informants humans respecte de la segmentació discursiva dels textos. Malauradament, l'escassetat de dades que hem obtingut i la disparitat d'aquestes dades han dificultat l'obtenció de mesures estadístiques significatives.

Per al disseny del nostre experiment hem pres com a model el treball de Passoneau i Litman (1993, 1997), un experiment amb jutges humans per tasques de segmentació de poca granularitat. Donat que volen recollir les intuicions dels humans, aquestes autores tan sols donen als informants un vague criteri de segmentació, basat en el reconeixement de les intencions dels parlants. Malgrat les divergències entre els diversos jutges, identifiquen dues tendències majoritàries: els qui segmenten molt, en segments petits, i els qui segmenten poc, en segments molt grans.

El nostre objectiu també era recollir les intuïcions dels parlants respecte dels segments tal com els hem definit en l'apartat 4.1.2., per la qual cosa no hem donat criteris extensos i inambigus als parlants, sinó que tan sols hem explicitat una vaga noció de segment i alguns exemples (vegeu en l'Annex III les instruccions per segmentació humana). La informació que esperàvem obtenir d'aquest experiment era observar com un humà dividia un text en segments similars als que identifica el segmentador. Això ens hauria de permetre validar el concepte de segment de què partíem i detectar possibles disparitats entre les intuïcions dels parlants i els resultats de l'eina que hem desenvolupat.

Com hem dit, els resultats d'aquest experiment han estat poc informatius, per dues raons principals: per una banda, hem disposat d'un nombre estadísticament insignificant de jutges humans (4), i, per altra banda, els resultats han estat molt diferents entre ells. Podem distingir tres tendències principals:

- dos informants amb formació lingüística universitària han segmentat el text seguint el que podrien semblar **criteris gramaticals**. Segons els seus propis comentaris, aquests informants van realitzar una anàlisi morfosintàctica del text per a segmentar-lo.
- un informant amb formació universitària no lingüística ha segmentat el text seguint el que podrien semblar criteris de **blocs d'informació**, més grans que la segmentació dels dos informants anteriors però sense entrar-hi en contradicció, és a dir: aquest informant no reconeixia fronteres de segment que no haguessin estat reconegudes pels dos informants anteriors, però reconeixia menys segments. No justificava les seves decisions de manera consistent, però era inflexible a l'hora de mantenir els límits de segment.
- un informant amb formació bàsica ha segmentat el text seguint un criteri que sembla de **longitud** de segment, tot i que no va justificar les seves decisions de manera consistent i tampoc mostrava seguretat en mantenir els límits de segment.

Malgrat l'escassetat de dades, no hem trobat que els parlants recorrin a cap concepte psicològic clar, d'acord amb la seva competència lingüística, per abordar la noció de segment en un nivell de granularitat alt, sinó que cada informant l'aborda d'acord amb algun coneixement adquirit. Si volem avaluar els resultats de la nostra eina de segmentació respecte de les intuïcions humanes caldrà, doncs, aprofundir en la definició de segment també des del punt de vista cognitiu, i dissenyar un nou experiment que pugui recollir-la de manera adequada.

No obstant, hem de dir que la noció de segment de baixa granularitat sembla poc natural per a un parlant. La segmentació és un procés necessari en l'arquitectura del sistema d'anàlisi retòrica dels textos, però no tenim cap raó per suposar que un humà processi el discurs de la mateixa manera que ho farà el sistema que hem proposat. És per això que també podem deixar l'avaluació amb criteris humans per a les tasques que tinguin un correlat de dimensió clarament humana, com ara la interpretació retòrica o el resum de textos. Quan aprofundim en la definició de segment haurem de determinar si l'avaluació humana resulta pertinent o no per a aquesta tasca de PLN.

4.3 Conclusions

En aquest capítol hem presentat una primera aproximació al concepte de segment discursiu en un nivell intraoracional. Això ens ha portat a abordar algunes qüestions certament interessants i complexes sobre les interaccions entre estructura sintàctica i estructura discursiva, com per exemple la divisió de l'oració en blocs caracteritzats sintàcticament però amb una funció discursiva clara o la relació entre segments que formen proposició i que no en formen i les dependències oracionals. No obstant, hem deixat de banda altres fenòmens que també tenen una gran rellevància i poder explicatiu en la interfície sintaxi-discurs, com la focalització, tematització i rematització, els patrons pronominals, les repeticions parcials, etc. Creiem que aquestes i altres qüestions requereixen un tractament detingut i aprofundit per tal d'assolir un nivell explicatiu satisfactori en el nivell oracional i textual.

Hem presentat també el desenvolupament d'una eina de segmentació automàtica que posa a la pràctica el concepte de segment que hem tractat i treu partit de la informació de marcadors del discurs que hem recollit i sistematitzat. Els resultats d'aquesta eina estan en procés de refinament. Intentarem incorporar progressivament claus que ens permetin identificar els segments i reduir al mínim el marge d'error, com per exemple:

- patrons regulars d'**interacció entre tipus de segments**, de manera que puguem determinar si un marcador ambigu és inambigu en una determinada seqüència de segments i/o marcadors[9], és a dir, ampliar el context que tenim en compte per determinar si un marcador és ambigu o inambigu

- patrons regulars d'**interacció entre segments i posicions oracionals** (inicial, final o intermitja), de manera que puguem ampliar les tendències que ja hem presentat més amunt sobre la relació entre , per exemple, que en posició inicial d'oració alguns segments

  exemple: *Los diputados socialistas Jesús Caldera y José Segura afirmaron ayer que todos los indicios demuestran que "se está produciendo un pelotazo", del que se benefician empresas del ministro de Asuntos Exteriores, Abel Matutes.*

  Com que el segmentador és conservador respecte de la detecció de segments, no reconeix darrera de segment cap tipus de segment que pugui ser subcategoritzat. És per això que aquesta oració seria analitzada de la següent manera:

  Los diputados socialistas Jesús Caldera y José Segura afirmaron ayer que todos los indicios demuestran que ***literal_*** *["se está produciendo un pelotazo"],* ***rel_*** *[**del que se benefician empresas del ministro de Asuntos Exteriores**], Abel Matutes.*

[9]Recollint la idea d''esquema' de la RST (vegeu apartat 2.2.2)

Tanmateix, si confirmem que els segments de tipus 'aposició' són inambigus en posició final d'oració i precedits de certs tipus de segment enniuat, com per exemple les frases de relatiu, podríem formular la següent regla de la gramàtica:

final ==> rel, coma, aposicio.

Amb aquesta regla, l'anàlisi de l'oració anterior seria:

Los diputados socialistas Jesús Caldera y José Segura afirmaron ayer que todos los indicios demuestran que "se está produciendo un pelotazo", ***rel_*** [***del que se benefician empresas del ministro de Asuntos Exteriores***], ***aposicio_*** [***Abel Matutes***].

- presència de **negació** en l'oració o en els segments circumdants
- **co-ocurrència de marcadors** amb altres marcadors o amb certs ítems lèxics
- el funcionament del segmentador també pot millorar si apliquem **diferents estratègies d'anàlisi**: per exemple, si no duem a terme una anàlisi linear, sinó que prenem com a base la plausabilitat de certs tipus de segment
- esperem que en el futur puguem incorporar al segmentador informació sobre les **entitats amb nom** (Arévalo i Simón 2000), que tenen uns patrons segmentals específics i, a més, poden contribuir a determinar els tipus de segment, com per exemple les aposicions (Alonso 2001).

Creiem que, a mesura que aprofundim en l'estudi dels textos, podrem trobar, sistematitzar i generalitzar aquest tipus d'evidències i tendències, per ampliar el tipus de fenòmens que el segmentador és capaç de reconèixer i tractar de manera fiable.

L'eina de segmentació que hem desenvolupat, malgrat que es troba en desenvolupament, pot constituir ja un bon ajut per a sistemes de PLN que requereixin una delimitació de blocs inferiors a l'oració, com per exemple certs tipus d'anàlisi sintàctica, tasques de resolució d'anàfora, etc. Pel que fa al resum automàtic, una eina d'aquestes característiques resulta útil per a sistemes basats en l'extracció de frases literals del text, ja que contribueix a reduir el volum de text extret, mantenint-ne la coherència. No oblidem, però, que l'objectiu principal d'aquesta eina és realitzar un pre-procés del text per a la seva interpretació retòrica, i que per tant els resultats estaran marcadament determinats per aquest objectiu.

Capítol 5

Conclusions

En aquest treball hem presentat un estudi preliminar per desenvolupar una aproximació al resum automàtic en llengua espanyola. Després de considerar diverses línies de treball en aquesta àrea, hem optat per explotar la informació superficial sobre un aspecte de l'estructura discursiva dels textos, l'anomenada 'estructura retòrica'. Hem considerat que aquest aspecte dels textos aporta informació útil per obtenir un resum automàtic satisfactori, tal com apunten Ono et al. (1994) i Marcu (1997), i, a més, està a l'abast de les possibilitats actuals del PLN per a l'espanyol. Hem proposat una línia de treball per desenvolupar un sistema de PLN que doni compte d'aquest tipus d'informació textual. Aquest sistema ha de contribuir a l'elaboració d'un resum automàtic de textos en espanyol.

Hem dut a terme un estudi dels fenòmens lingüístics que poden facilitar el processament computacional de l'estructura retòrica dels textos. En primer lloc, hem **delimitat el nostre objecte d'estudi**: hem definit les unitats de treball 'marcador del discurs' i 'segment'. Un cop fet això, hem dut a terme un **estudi sobre corpus** per tal de detectar la informació necessària per al processament automàtic d'aquesta informació discursiva. A partir d'aquest estudi, hem **sistematitzat** aquesta informació per al seu ús computacional, desenvolupant dues fonts de coneixement lingüístic per a l'abordatge de l'anàlisi retòrica dels textos: una **lexicó computacional de marcadors del discurs** per a l'espanyol i una **gramàtica de segmentació** retòrica. Finalment, hem desenvolupat una **aplicació informàtica** que suposa un primer pas al processament de l'estructura retòrica dels textos: la seva segmentació en unitats discursives mínimes.

Treball futur

Hem encetat doncs una línia de treball per abordar el resum automàtic en espanyol. El principal treball futur consistirà a desenvolupar els recursos i eines de PLN necessaris per donar compte l'estructura retòrica dels textos, amb l'objectiu final d'aconseguir un sistema de resum automàtic o bé un mòdul d'ajut a un sistema de resum en què col·laborin diversos tipus d'informació sobre els textos. Caldrà, doncs, aprofundir el treball que hem iniciat aquí pel que fa a marcadors del discurs i segmentació, i caldrà desenvolupar el mòdul d'interpretació retòrica dels textos i també els procediments necessaris per elaborar un resum partir de la informació textual de què disposem: retòrica i d'altres aspectes dels textos.

Malgrat que el treball que proposem aquí vol ser independent de gènere i domini conceptual, també pot resultar interessant de desenvolupar alguna aplicació dependent de domini, per tal de comprovar si l'anàlisi d'estructura retòrica és més efectiu quan una de les variables amb què es treballa habitualment queda eliminada. Un experiment d'aquest tipus pot ser molt revelador pel que fa a l'abast dels mecanismes lingüístics retòrics.

A mesura que hem aprofundit en aquest treball, hem detectat alguns punts complexos o conflictius que caldrà resoldre per tal d'assolir un nivell satisfactori d'explicació dels fenòmens discursius que tractem i, com a conseqüència, un tractament computacional adequat dels textos. Sovint, la solució d'aquests punts conflictius passa per una estreta col·laboració amb àrees properes de la lingüística i del PLN.

En el nostre treball hem considerat oportú d'identificar elements discursius en un nivell intraoracional i treballant amb informació morfosintàctica. En situar-nos en aquest nivell del llenguatge, hem trobat que les **interaccions entre l'estructura sintàctica i l'estructura discursiva** dels textos poden ser molt informatives. No obstant, els fenòmens sintàctico-discursius han estat sistemàticament oblidats pel PLN (amb la notable excepció de Polanyi (1988, 1995)), però, en canvi, desperten un gran interès en lingüística teòrica (Givón 1979). Una de les línies de recerca futura consistirà a aprofundir en les interaccions entre aquests dos nivells lingüístics, per tal d'avançar en l'aproximació a una explicació satisfactòria de qüestions com

– la interacció entre **blocs sintàctics** i **blocs discursius**: ja hem presentat com les posicions oracionals determinen els patrons segmentals més característics (apartat 4.2.2), igualment, els patrons segmentals poden contribuir a delimitar l'abast de certes estructures sintàctiques

– les **dependències** entre constituents oracionals: ja hem assenyalat com els elements que no pertanyen a la configuració prototípica d'un event solen ser els que estableixen una relació que podem considerar retòrica amb el nucli oracional. En canvi, aquest nucli oracional estaria consituït d'elements imprescindibles per a la configuració d'un event. La qüestió de la subcategorització dels constiutents oracionals és central per a la lingüística teòrica i també la computacional. També per obtenir l'estructura retòrica d'un text hem de recórrer a aquesta noció. No obstant, hem observat que paràmetres com el pes lèxic dels marcadors del discurs contribueixen a aclarir en certa mesura aquesta qüestió. Aquesta sembla, doncs, una qüestió que pot aportar informació interssant sobre les interaccions entre el nivell sintàctic i el discursiu, alhora que pot contribuir a solucionar problemes de PLN en ambdós nivells lingüístics.

– les **funcions oracional i discursiva dels marcadors** del discurs: algunes marques morfosintàctiques, com el temps i mode verbal, i la presència de certs adverbis o formes pronominals poden contribuir a determinar quan un marcador del discurs amb doble funció està assenyalant una relació retòrica entre segments.

– el **contingut retòric dels marcadors polisèmics**: igualment, marques morfosintàctiques i la presència de certs ítems lèxics poden contribuir a desambiguar el contingut de certs marcadors

Per altra banda, creiem que la representació dels textos en qualsevol dels seus aspectes s'enriqueix considerablement si s'incorporen altres tipus d'informació textual. Com ja hem apuntat abans, podem considerar un text com un conjunt d'informacions heterogènies lligades entre elles de manera que arriben a configurar un tot cohesionat. Així doncs, no n'hi ha prou de donar compte de les diverses facetes d'un text: caldrà, igualment, treballar les **interaccions entre tipus diversos d'informació textual**. Hem trobat, per exemple, que la detecció i seguiment de les entitats d'un event contribueix en gran mesura a assenyalar de manera clara el fil narratiu d'un text, i així determinar la nuclearitat d'alguns segments en relacions retòriques poc definides. Igualment, les entitats del discurs presenten patrons segmentals típics que seria convenient de sistematitzar i explotar per millorar ela segmentació automàtica dels textos. També pot resultar convenient tractar de manera col·laborativa les informacions obtingudes d'altres aproximacions al resum automàtic, com el mètode de cadenes lèxiques (Fuertes 2001), la posició de les paraules i oracions en un document, les paraules clau de contingut, etc.

Per últim, i no per això menys important, caldrà dissenyar i aplicar un protocol d'**avaluació** del funcionament del sistema que volem desenvolupar. Com ja hem dit, avaluar els sistemes de resum automàtic no és una tasca fàcil, ja que hi ha un factor subjectiu important. Els resums no es valoren tan sols per la seva informativitat objectiva, sinó també per la seva qualitat com a textos coherents i cohesius, cosa certament difícil d'avaluar de manera objectiva. No obstant, creiem que cal dur a terme una avaluació tan rigorosa com sigui possible del sistema, comptant sempre que sigui possible amb l'opinió de jutges aliens al projecte. Així podrem detectar les mancances i limitacions del nostre treball, i podrem millorar-lo i adequar-lo a les necessitats reals dels usuaris.

Aquest haurà de ser, doncs, el marc del nostre treball de tesi doctoral. Tractarem de mantenir-nos oberts al desenvolupament de les tècniques de PLN, per incorporar aquells avenços que suposin una millora en la nostra aproximació al tractament dels textos. Igualment, no perdrem de vista els progressos en l'àrea de RA, per tal de mantenir en continua revisió la nostra aproximació i redirigir, si cal, el nostre treball.

Agraïments

Aquest treball es du a terme en el marc del projecte del Ministeri d'Educació i Cultura X-TRACT, nº PB98-1226, amb una beca de Formació de Personal Investigador del Ministeri de Ciència i Teconologia. També comptem amb el suport del projecte HERMES, TIC 2000-0335-C032-02, i del projecte LEXESP-III, APC 99-0105 de la DGICYT.

Bibliografia

[1] ALCINA FRANCH, J., BLECUA, J. M., (1975), *Gramática española*, Ariel, Barcelona

[2] ALONSO ALEMANY, L. (2001), "A proposal for an Automated Text Summarisation System: Integrating Rhetorical Structure Theory and Entity Detection and Tracking", en CHAUDIRON, S., FLUHR, C., *3ème Congrès du Chapitre français de l'ISKO: Filtrage et résumé automatique de l'information sur les réseaux*, Universitat de Nanterre - Paris X

[3] AMITHAY, E., K. RICHMOND, A. SMITH, (1997), "Detecting Subject Boundaries within Text: A Language Independent Statistical Approach", *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pp. 4754

[4] ANDO, R. K., BOGURAEV, B. K., BYRD, R. J., NEFF, M. S., (2000), "Multi-document summarization by visualizing topical content", en *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization*

[5] ANSCOMBRE, J. C., DUCROT, O., (1994), *La argumentación en la lengua*, Madrid, Gredos (versión original: *L'argumentation dans la langue*, Bruxelles, Mardaga, 1983)

[6] ARÉVALO, M., ALONSO, L., TAULÉ, M., MARTÍ, M.A., (2001), *Documentación sobre el Analizador Morfológico para el Castellano (AMCAS)*, X-Tract 01/01 Working Paper, CLiC (Centre de Llenguatge i Computació), Universitat de Barcelona, Barcelona

[7] ATSERIAS, J., CASTELLÓN, I., CIVIT, M., (1998a), "Syntactic Parsing of Unrestricted Spanish Text", *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada

[8] ATSERIAS, J.; CARMONA, J.; CASTELLÓN, I.; CERVELL, S.; CIVIT, M.; MÀRQUEZ, L.; MARTÍ, M.A.; PADRÓ, L.; PLACER, R.; RODRÍGUEZ, H.; TAULÉ, M.; TURMO, J.; (1998b), "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text", *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada

[9] S. AZZAM, K. HUMPHREYS, R. GAIZAUSKAS, (1999), "Using Coreference Chains for Text Summarization", *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, Baltimore

[10] BARZILAY, R., ELHADAD, M., (1997), "Using lexical chains for text summarization", *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, ACL, Madrid

[11] BARZILAY, R., MCKEOWN, K., ELHADAD, M., (1999), "Information Fusion in the Context of Multi-Document Summarization", *Proceedings of the ACL'99*, pp. 550–557

[12] BEEFERMAN, D., BERGER, A., LAFFERTY, J., (1999), "Statistical models for text segmentation", *Machine Learning*, nº 34, vol. 1-3, pp. 177-210

[13] BELLO, A., (1847), *Gramática de la lengua castellana destinada al uso de los americanos*, (edició de de Ramón Trujillo, Madrid, Arco/Libros, 1988)

[14] Bloomfield, L, (1933), *Language*, Holt, Rinehard and Winston, Nova York

[15] Boguraev, B., Neff, M., (2000a), "Discourse segmentation in aid of document summarization", *Proceedings of Hawaii International Conference on System Sciences (HICSS-33), Minitrack on Digital Documents Understanding*, Maui, Hawaii

[16] Boguraev, B., Neff, M., (2000b), "Lexical Cohesion, Discourse Segmentation and DocumentSummarization", *RIAO-2000*, París

[17] Bosque, I., Demonte, V., (eds.), (1999), *Gramática Descriptiva de la Lengua Española*, Espasa Calpe, Madrid

[18] Bosque, I., (2000), pròleg a Montolío Durán, E., (2001), *Conectores de la lengua escrita*, Barcelona, Ariel

[19] Brandow, R., Mitze, K. , Rau, L., (1994), "Automatic condensation of eletronic publications by sentence selection", *Information Processing and Management*, nº 31, vol. 5, pp. 675-685

[20] Briz, A., (1993a), "Los conectores pragmáticos en español coloquial (I). Su papel argumentativo", *Contextos*, XI:21-22, pp. 369-399

[21] Briz, A., (1993a), "Los conectores pragmáticos en español coloquial (II). Su papel metadiscursivo", *EAc*, nº 59, pp. 39-56

[22] Carberry, S., (1990), *Plan Recognition in Natural Language Dialogue*, MIT Press

[23] Carbonell, J., Goldstein, J., (1998), "The use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", poster, *SIGIR'98*, Melbourne,

[24] Carmona, J., S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, J. Turmo, (1998), "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text", en *Proceedings of First International Conference on Language Resources and Evaluation (LREC'98),* Granada

[25] Casado Velarde, M., (1991), "Los operadores discursivos *es decir, esto es, o sea* y *a saber* en español actual: valores de lengua y funciones textules", *LEA*, nº 13, pp. 87-116

[26] Chali, Y., Matwin, S., Szpakowicz, S., (1999), "Querybiased text summarization as a question-answering technique", en *Proceedings of the AAAI Fall Symposium on Question Answering Systems*, pp. 52-56

[27] Civit, M., Castellón, I., Martí, M. A., Taulé, M., (2000), "LEXPIR: A verb lexicon for Spanish", *Cuadernos de Filología Inglesa,* vol. 9.1, Publicaciones de la Universidad de Granada

[28] Civit, M., Castellón, I., (1998), "Gramesp: una gramática de corpus para el español", *Revista de AESLA*, La Rioja

[29] Cortés Rodríguez, L., (1995a), "Bibliografía: marcadores del discurso (I)", *EAc*, nº 63, pp. 63-82

[30] Cortés Rodríguez, L., (1995b), "Bibliografía: marcadores del discurso (y II)", *EAc*, nº 64, pp. 75-94

[31] De Beaugrande, R. A., Dressler, W. U., (1981), *Introduction to Text Linguistics*, Longman. (en espanyol: *Introducción a la Lingüística del texto*, Ariel, Barcelona, 1997)

[32] DeJong, G., (1978), *Fast Skimming of News Stories: the FRUMP System*, Tesi doctoral, Universitat de Yale

[33] Edmunson, H. P., (1969), "New methods in automatic extracting", *Journal of the ACM*, nº 16, vol. 2, pp. 264-285

[34] Endres-Niggemeyer, B., (1998), *Summarizing Information*, Springer, Berlin

[35] Fuentes Fort, M., (2001), *Lexical Chains: Comparative Study, Implementation and Applications Areas*, report de recerca IMA 01-01-RR, Universitat de Girona

[36] Fuentes Rodríguez, C., (1987), *Enlaces extraoracionales*, Sevilla, Alfar

[37] Garcés, G., (1791), *Fundamento del vigor y elegancia de la lengua castellana, expuesto en el propio y vario uso de sus partículas,* (edició de M. Ballesteros, Rivadeneyra, Madrid, 1852)

[38] Gili Gaya, S., (1943), *Curso superior de sintaxis española*, (Barcelona, Bibliograf, 1961)

[39] Givón, T. (ed.), (1979), *Discourse and Syntax, Syntax and Semantics,* vol. 12, Academic Press, New York.

[40] Goldstein, J., Kantrowitz, M., Mittal, V. O., Carbonell, J., (1999), "Summarizing Text Documents: Sentence Selection and Evaluation Metrics", en *Proceedings of SIGIR-99*, Berkeley

[41] González, A. L., Goñi, J. M., González, J., C., (1995), "Un analizador morfológico para el castellano basado en chart", en *Actas de la VI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'95)*, Alicante

[42] Gonzalo, J., F. Verdejo, I. Chugur, J. Cigarrán, (1998), "Indexing with WordNet synsets can improve Text Retrieval", en *Proceedings of the COLING/ACL'98 Workshop on usage of Wordnet for NLP applications.*

[43] Gonzalo, J., A. Peñas, F. Verdejo, (1999), "Lexical ambiguity and Information Retrieval Revisited", en *Proceedings EMNLP/VLC'99*

[44] Grice, H. P., (1975), "Logic in conversation", *Syntax and semantics,* vol. III: Speech Acts, Peter Cole i Jerry L. Morgan (eds.), , Academic Press, Nova York, pp. 41-58

[45] Grosz, B. J., Joshi, A. K., Weinstein, S., (1995), "Centering: A framework for modeling the local coherence of discourse", *Computational Linguistics*, nº 21, vol. 2, pp. 203-226

[46] Grosz, B. J., Sidner, C. L., (1986), "Attention, intention, and the structure of discourse", *Computational Linguistics*, nº 12, vol. 3, pp. 175-204

[47] Hahn, U., Strube, M., (1997), "Centering in-the-large: Computing referential discourse segments", *Proceedings of the 35 th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, pp. 104-111

[48] Hajime, M., Takeo, H., Manabu, O., (1998), "Text Segmentation with Multiple Surface Linguistic Cues", *Proceedings of the COLING-ACL'98*, pp. 881-885

[49] Halliday, M.A.K., Hasan, R., (1976), *Cohesion in English*, Longman

[50] Hearst, M. A., (1994), "Multi-paragraph segmentation of expository text", *Proceedings of the ACL*

[51] Heinonen, O., (1998), "Optimal multi-paragraph text segmentation by dynamic programming", *Proceedings of COLING-ACL'98*

[52] Hirschberg J., Litman, D., (1993), "Empirical Studies on the Disambiguation of Cue Phrases", *Computational Linguistics,* nº 19, pp. 501-530

[53] HOBBS, J. R., (1978), "Why is discourse coherent", nota tècnica 176, SRI International, Artificial Intelligence Center

[54] HOBBS, J. R., (1985), *On the coherence and structure of discourse*, Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford

[55] HOVY, E., (1993), "Automated Discourse Generation Using Discourse Structure Relations", *Artificial Intelligence*, nº 63, pp. 341-385

[56] HOVY, E., LIN, C. Y., (1997), "Automated Text Summarization in SUMMARIST", en MAYBURY I MANI, (1997), (eds.), *Intelligent Scalable Text Summarization*

[57] HUMPHREYS, K., GAIZAUSKAS, R., AZZAM, S., HUYCK, C., MITCHELL, B., CUNNINGHAM, H., (1998), "University of Sheffield: description of the LaSIE-II System as used for MUC-7", en *Proceedings of the Seventh Message Understanding Conference*

[58] MAIER, E., HOVY, E.H., (1991), "A Metafunctionally Motivated Taxonomy for Discourse Structure Relations", *Third European Workshop on Natural Language Generation*

[59] HARMAN, D., VOORHEES, E. M., (1996), *The fifth Text REtrieval Conference (TREC-5)*, National Institute of Standards and Technology NIST SP 500-238

[60] HOVY, E.H., MAIER, E., (1997), "Parsimonious and Profligate: How Many and Which Discourse Structure Relations?", *Discourse Processes*

[61] JING, H., (1998), *Summary Generation through Intelligent Cutting and Pasting of the Input Document*, proposta de tesi doctoral, Department of Computer Science, University of Columbia

[62] JING, H., BARZILAY, R., MCKEOWN, K., ELHADAD, M., (1998), "Summarization evaluation methods experiments and analysis", en *AAAI Intelligent Text Summarization Workshop*, pp. 60-68

[63] KAN, M.-Y., KLAVANS, J. L., MCKEOWN, K. R., (1998), "Linear segmentation and segment significance", *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98)*, pp. 197-205, Montreal, Canada

[64] KAN, M.-Y., MCKEOWN, K. R., (1999), *Information Extraction and Summarization: Domain Independence through Focus Types*, report tècnic, Computer Science Department, Columbia University, New York

[65] KNOTT, A., DALE, R., (1994), "Using linguistic phenomena to motivate a set of coherence relations", *Discourse Processes*, vol. 18, nº 1, pp. 35-62

[66] KNOTT, A., (1995), *A Data-Driven Methodology for Motivating a Set of Coherence Relations*, tesi doctoral, University of Edinburgh, Edimburg

[67] KRUIJ -KORBAYOVA, I., WEBBER, B., (2000), "Discourse connectives, inference and information structure", en BOS, J. AND KOHLHASE, M., (eds.), *Proceedings of the ICoS-2 Workshop on Inference in Computational Semantics*, Dagstuhl, Alemanya

[68] KUHN, T., (1962), *The Structure of Scientic Revolutions*, University of Chicago Press, Chicago

[69] KUPIEC, J., PEDERSEN, J., CHEN, F., (1995), "A Trainable Document Summarizer", en *Proceedings of the 18th ACM-SIGIR Conference*, pp. 68-73

[70] LEHMAM, A., (1995), *Le résumé automatique des textes scientifiques et techniques; aspects linguistiques et computationnels: réalisation d'un prototype procédant par extraction de phrases du texte source -RAFI- (Résumé Automatique par Fragments Indicateurs)*, tesi doctoral, Universitat Nancy 2

[71] LEHMAM, A., BOUVET, P., (2001), "Évaluation, rectification et pertinence du résumé de texte automatique pour une utilisation en réseau", en CHAUDIRON, S., FLUHR, C., *3ème Congrès du Chapitre français de l'ISKO: Filtrage et résumé automatique de l'information sur les réseaux*, Universitat de Nanterre - Paris X

[72] LIN, C.-Y., HOVY, E., (1997), "Identifying Topics by Position", en *Proceedings of the Applied Natural Language Processing Conference (ANLP-97)*, 283-290, Washington

[73] LITMAN, D., PASSONNEAU, R., (1993), *Empirical evidence for intention-based discourse segmentation*, MIT Press, Cambridge, Massachussets

[74] LLITERAS, M., (1995), "El *Tratado de las partículas* de G. Garcés (1791) y su presencia en A. Bello", en M. T. ECHENIQUE ET AL. (eds.), *Historia de la lengua española en América y España*, Universitat de València, València, pp. 371-381

[75] LUHN, H. P., (1958), "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, 159-165

[76] MANI, I., BLOEDORN, E., (1997), "Multidocument summarization by graph search and matching", en *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-97)*, pp. 622-628

[77] MANI, I., BLOEDORN, E., (1998), "Machine Learning of Generic and User-focused Summarization", en *Proceedings of the 15th National Conference on Artificial Intelligence*, pp. 821-826

[78] MANI, I., HOUSE, D., KLEIN, G., HIRSCHMAN, L., OBRST, L., FIRMIN, T., CHRZANOWSKI, M., SUNDHEIM, B., (1998), *The tipster SUMMAC text summarization evaluation: Final report*, report tècnic, DARPA, EUA

[79] MANI, I., MAYBURY, M., (1999), *Advances in Automatic Text Summarization*, MIT Press

[80] MANN, W. C., THOMPSON, S. A., (1988), "Rhetorical Structure Theory: Towards a functional theory of text organization", *TEXT*, nº 8, vol. 3, pp. 243-281.

[81] MARCU, D., (1997), "The Rhetorical Parsing of Unrestricted Natural Language Texts", *Proceedings of the 35th Annual Meeting of the Association Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, pp. 96-103

[82] MARCU, D., (1998), "To build text summaries of high quality, nuclearity is not sufficient", en *The Working Notes of the the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1-8

[83] MARCU, D., (1999), "The automatic construction of large-scale corpora for summarization research", en *The 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 137-144

[84] MÁRQUEZ, L., PADRÓ, L., (1997), "A Flexible POS Tagger Using an Automatically Acquired Language Model", *Proceedings of EACL/ACL,* Madrid

[85] MARTÍN ZORRAQUINO, M. A., MONTOLÍO, E., (coords.), (1998), *Los Marcadores del Discurso: teoría y análisis*, Arco Libros, Madrid

[86] MARTÍN ZORRAQUINO, M. A., PORTOLÉS, J., (1999), "Los marcadores del discurso", en I. Bosque i V. Demonte, (eds.), *Gramática Descriptiva de la Lengua Española*, vol. III, Espasa Calpe, Madrid, pp. 4051-4213

[87] Martínez Unanue, R., (1999), *Alineación Automática de Corpus Paralelos: Una Propuesta Metodológica y su Aplicación a un Dominio de Especialidad*, tesi doctoral, Universidad de Deusto, Bilbao

[88] McKeown, K., Radev, D. R., (2000), "Generating Summaries of Multiple News Articles", *Proceedings of the International Conference on Computational Linguistics*

[89] Mederos, H. (1988), *Procedimientos de cohesión en el español actual*, Tenerife, Cabildo Insular de Canarias

[90] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., (1990), *Five Papers on WordNet*, report tècnic, nº 43, Cognitive Science Laboratory, Princeton University

[91] Monedero, J., González, J. C., Goñi, J. M., Iglesias, C. A., Nieto, A., (1995), "Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS", en *Actas del XI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'95)*, pp. 241-254, Bilbao

[92] Montolío Durán, E., (1998), "La Teoría de la Relevancia y el estudio de los marcadores discursivos", en M. A. Martín Zorraquino i E. Montolío Durán (coords.,)*, Los Marcadores del Discurso: teoría y análisis*, Arco Libros, Madrid, pp. 93-120

[93] Montolío Durán, E., (2000), "La conexión en el texto escrito académico. Los conectores", en E. Montolío (coord.), *Manual práctico de escritura académica*, vol. II, Barcelona, Ariel, pp. 105-164

[94] Montolío Durán, E., (2001), *Conectores de la lengua escrita*, Barcelona, Ariel

[95] Moore, J. D., Pollack, M. E., (1992), "A problem for RST: The need for multi-level discourse analysis", *Computational Linguistics*, nº 18, vol. 4, pp. 537-544

[96] EAGLES, (2000), "Encoding conventions for segmentation and grammatical annotation", en *Corpus Encoding Standard,* Document CES 1. Version 1.5, http://www.cs.vassar.edu/CES/

[97] Nakao, (2000), "An Algorithm for One-Page Summarization of a Long Text Based on Thematic Hierarchy Detection", *ACL Proceedings*

[98] Nomoto, T., Nitta, Y., (1993), "Resolving Zero Anaphora in Japanese", *ACL Proceedings or Sixth European Conference*, pp. 315-321

[99] Ono, K., Sumita, K., Miike, S., (1994), "Abstract generation based on rhetorical structure extraction", en *Proceedings of International Conference on Computational Linguistics*, pp. 344-348

[100] Padró, L., (1997), *A Hybrid Environment for Syntax-Semantic Tagging*, tesi doctoral, Departament de Lenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona

[101] Paice, C. D., (1990), "Constructing Literature Abstracts by Computer: Techniques and Prospects", en *Information Processing & Management*, nº 26, vol. 1, pp. 171-186

[102] Passonneau, R. J., Litman, D. J., (1993), "Intention-based segmentation: Human reliability and correlation with linguistic cues", *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* , pp. 148-155

[103] Passonneau, R. J., Litman, D. J.,, (1997), "Discourse segmentation by human and automated means", *Computational Linguistics*, nº 23, vol. 1, pp. 103-140

[104] Poesio, M., (1991), "Expectation-based recognition of discourse segmentation", *Proceedings of the AAAI Fall Symposium on Discourse Structure*

[105] Polanyi, L., (1988), "A formal model of the structure of discourse", *Journal of Pragmatics*, 12

[106] Polanyi, L., (1995), *The Linguistic Structure of Discourse.* Centre for the Study of Language and Information, Stanford

[107] Pons Bordería, S., (1998), *Conexión y conectores: estudio de su relación en el registro informal de la lengua*, Anejo XXVII de Cuadernos de Filología, Universitat de València

[108] Portolés, J., (1998a),"Teoría de la argumentación en la lengua y los marcadores del discurso", en M. A. Martín Zorraquino i E. Montolío Durán, (eds.), pp. 243-264

[109] Portolés, J., (1998b), *Marcadores del discurso*, Ariel, Barcelona

[110] Radev, D., (1996), "An architecture for distributed natural language summarization", en *Proceedings of the 8th International Workshop on Natural Language Generation: Demonstrations and Posters*, pp 45-48

[111] Radev, D., (2000a), *Text Summarization*, tutorial en ACM-SIGIR, Atenes

[112] Radev, D., Jing, H., Budzikowska, M., (2000b), "Summarization of multiple documents: clustering, sentence extraction, and evaluation", en *ANLP/NAACL Workshop on Summarization*, Seattle

[113] Radev, D., McKeown, K., (1998), "Generating natural language summariesfrom multiple online sources", *Computational Linguistics*, nº 24, vol. 3, 469-501

[114] Redeker, G., (1990), "Ideational and pragmatic markers of discourse structure", *Journal of Pragmatics*, nº 14, pp. 367-381

[115] RILE, (2000), *RILE, Servidor de Recursos para el desarrollo de la Ingeniería Lingüística en Español*, http://rile.sema.es/index.html

[116] Rodríguez, H., (1989), *GUAI: un generador automático de interfaces en lengua natural*, tesi doctoral, LSI-UPC

[117] Salinas, M. (1541), *Retórica en lengua castellana*, Madrid, Editora Nacional, 1980

[118] Ted J.M. Sanders, Wilbert P.M. Spooren, Leo G.M. Noordman, (1993), "Coherence relations in a cognitive theory of discourse representation, *Cognitive Linguistics*, nº 4, vol. 2, pp. 93–133

[119] O.Santana, J.Perez, F.Carreras, G.Rodriguez, Z.Hernández, J.Duque., (2000), *EMERGE. Flexionador y Lematizador de palabras del español*, http://protos.dis.ulpgc.es/morfolog/morfolog.htm, Universidad de Las Palmas de Gran Canaria,

[120] Santos, D., (1999) "Toward Language-dependent Applications", *Machine Translation*, nº 14, pp. 83-112

[121] Scha, R., Polanyi, L., (1988), "An augmented contextfree grammar for discuorse", en *Proceedings of the 12th International Conference on Computational Linguistics, COLING-88*, pp 22-27

[122] Schank, R., Abelson, R., (1977), *Scripts, Plans, Goals, and Understanding*, Hillsdale, NJ: Lawrence Erlbaum

[123] Schauer, H., Hahn, U, (2000), "Phrases as carriers of coherence relations", en L.R. Gleitman i A.K. Joshi (eds.), *CogSci 2000 – Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, University of Pennsylvania, Philadelphia, USA, pp. 13-15

[124] SCHIFFRIN, D., (1987), *Discourse Markers*, Cambridge University Press, Cambridge

[125] SHANNON, C.E., (1951), "Prediction and entropy of printed English", *Bell System Technical Journal*, pp. 50-64

[126] SPARCK JONES, K., (1993), *Discourse Modelling for Automatic Summarising*, Report Tècnic nº 290, University of Cambridge

[127] STAIRMAND, M. A., (1996), *A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval*, tesi doctoral, Center for Computational Linguistics, UMIST, Manchester

[128] SUZUKI, Y., FUKUMOTO, F., SEKIGUCHI, Y., (1996), "Discourse segmentation for radio news", *ASA and ASJ Third Joint Meeting*, pp. 1009-1014.

[129] SWERTS, M., (1995), "Combining statistical and phonetic analyses of spontaneous discourse segmentation", *Proceedings of the 12th International Congress of Phonetic Sciences*, Estocolm, vol. 4, pp. 208-211

[130] VAN DIJK, T. A., KINTSCH, W., (1978), "Cognitive Psychology and Discourse: Recalling and Summarizing Stories", en W. U. DRESSLER, (ed.), *Current Trends in Textlinguistics*, vol. 2 de *Research in Text Theory*, pp. 61-80, de Gruyter, Berlin

[131] VAN DIJK, T.A. (1979), *Macrostructures: an interdisciplinary study of global structures in discourse, interaction, and cognition*, Hillsdale, N.J.: Erlbaum

[132] VOSSEN, P., P. DIEZ-ORZAS, W. PETERS, (1997), "The Multilingual Design of EuroWordNet", en P. VOSSEN, N.CALZOLARI, G. ADRIAENS, A. SANFILIPPO, Y. WILKS (eds.) *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid

[133] WALKER, M., WHITTAKER, S, (1990), "Mixed initiative in dialogue: An investigation into discourse segmentation", *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 70-78

[134] WASSON, M. (1998), "Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications", en *Proceedings of COLING-ACL '98 Conference*

[135] WEBBER, B. L., KNOTT, A., JOSHI, A. K. (1999), "Multiple discourse connectives in a lexicalised grammar for discourse", en BUNT, H., I THIJSSE, E., (eds.), *Proceedings from the 3rd International Workshop on Computational Semantics, IWCS-3*, Tilburg, pp. 309–325

[136] WIEBE, J. M., BRUCE, R. (1995), "Probabilistic classifiers for tracking point of view", *Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation*

[137] WING, C. S., SCHOLNICK, E. K., (1981), "Children's comprehension of pragmatic concepts expressed in *because, although, if*, and *unless*", *Journal of Child Language*, nº 8, pp. 347-365

[138] YAARI, Y., (1997), "Segmentation of expository texts by hierarchical agglomerative clustering", *Proceedings of RANLP*, Bulgària, pp. 59-65

[139] YUSHKIAVITSHUS, H., (1998), Discurs en la Ceremònia d'Obertura de la XXXIII Conferència Internacional de la Round Table on Archives (CITRA), Estocolm

[140] ZAVREL, J., DAELEMANS, W., (1999), "Recent Advances in Memory-Based Part-of-Speech Tagging", en *Actas del VI Simposio Internacional de Comunicacion Social*, Santiago de Cuba, pp. 590-597

Apèndix A

Annex I: Gramàtica de Segmentació

Tot seguit descrivim la gramàtica en què es basa el segmentador retòric que hem desenvolupat. Es tracta d'una gramàtica independent de context, implementada en l'analitzador TACAT (Atserias et al. 1998). El seu objectiu és la segmentació d'un text en llengua espanyola en unitats sintàcticament completes iguals o menors a l'oració, que anomenem segments (vegeu la definició de segment, apartat 4.1.2.). Per fer-ho utilitza informació morfosintàctica i de marcadors del discurs. Aquesta gramàtica ha estat construida amb l'objectiu de prioritzar la precisió per sobre de la cobertura, és a dir: reconeix menys segments que un humà podria reconèixer en el text, però els segments que reconeix són efectivament segments. Per això, tan sols utilitza la informació morfosintàctica i de marcadors del discurs quan aquesta informació és inambigua: en contextos concrets de pausa de puntuació, en presència de certes preposicions, etc. Això, però, no afecta la seva robustesa, ja que, independentment de la complexitat del text, sempre s'obté alguna anàlisi.

Aquesta gramàtica consta de 123 regles de composició de segments. El tractament del lèxic es basa en el que fa la gramàtica de Civit i Castellón (1998), i afegeix informació sobre marcadors del discurs en el seu aspecte de delimitació (vegeu apartat 3.2.2.). La unitat d'anàlisi més gran és la de paràgraf.

1. paragraf ==> ors, par.

   La unitat 'paràgraf' és útil per a l'anàlisi retòrica dels textos, ja que contribueix a limitar l'abast d'alguns marcadors. Per identificar-la, es realitza un pre-procés a l'anàlisi segmental, en què es substitueixen els salts de línia per una marca inambigua (#) que està descrita en el lèxic de la gramàtica com a marca de final de paràgraf.

2. ors ==> oracio, ors.

   Introdueix la recursivitat en el contingut dels paràgrafs: un paràgraf pot contenir un nombre indeterminat d'oracions.

3. ors ==> oracio.

   Restringeix el contingut dels paràgrafs: tot paràgraf ha de contenir, com a mínim, una oració. També limita la recursivitat de la regla anterior.

4. oracio ==> inicial, coma, migs, final, punt.

   Les regles 4 a 7 defineixen les configuracions possibles d'una oració. El contingut màxim possible (regla 4) és un segment amb format que pot estar en inici d'oració, seguit de text i/o segments amb format de manera recursiva, i un segment amb format que pot estar en final d'oració. El punt constitueix la marca dels límits de les oracions, la coma, la marca de límit d'aquestes tres posicions possibles de segments. D'aquestes tres posicions de segment, tan sols la central és obligatòria perquè es constitueixi una oració. Les regles següents (5 a 7) donen les variants sobre la opcionalitat de les altres dues posicions: la regla 5 defineix una oració sense segment amb format al final, la regla 6 defineix una oració sense segment amb format a l'inici, i la regla 7 defineix el contingut mínim possible d'una oració: un segment mig, sense segments amb format ni a l'inici ni al final.

***exemple*:** oracio(inicial(*Después de tratar de disimularlo*), migs(mig(text(*Sonia tuvo que admitir que el pastel*), format(rel(*que habíamos hecho entre todos con mucho esfuerzo*))), mig(text(*no le había gustado*))), final(*pero nos lo agradeció en el alma*).)

5. oracio ==> inicial, coma, migs, punt.

***exemple*:** oracio(inicial(*Después de tratar de disimularlo*), migs(mig(text(*Sonia tuvo que admitir que el pastel*), format(rel(*que habíamos hecho entre todos con mucho esfuerzo*))), mig(text(*no le había gustado*))).)

6. oracio ==> migs, final, punt.

***exemple*:** oracio(migs(mig(text(*Sonia tuvo que admitir que el pastel*), format(rel(*que habíamos hecho entre todos con mucho esfuerzo*))), mig(text(*no le había gustado*))), final(*pero nos lo agradeció en el alma*).)

7. oracio ==> migs, punt.

***exemple*:** oracio(migs(mig(text(*Sonia tuvo que admitir que el pastel*), format(rel(*que habíamos hecho entre todos con mucho esfuerzo*))), mig(text(*no le había gustado*))).)

8. migs ==> mig, migs.

Dóna compte de la recursivitat de segments en la posició central d'oració. Els límits de segment, comes i marques del tipus 'delfort', estan definits en les regles que caracteritzen cada tipus de segment.

9. migs ==> mig.

Restringeix la recursivitat en la posicio central d'oració: en aquesta posició ha d'haver-hi com a mínim una unitat de tipus 'mig'.

10. mig ==> text, coma, format, limit.

Defineix com ha d'ocórrer un segment amb format en posició intermitja d'oració (no final ni inicial): el límit inicial d'aquest tipus de segment és una coma, ha d'anar precedit necessàriament de text i seguit o bé de text, o bé d'una unitat del tipus 'limit': una coma, un segment de tipus 'fort'o o un segment de tipus 'feble'.

***exemple*:** oracio(migs(mig(text(*Sonia tuvo que admitir que el pastel*), format(rel(*que habíamos hecho entre todos con mucho esfuerzo*))), mig(text(*no le había gustado*))).)

11. mig ==> fort, limit.

Defineix el context d'ocorrença d'un segment del tipus 'fort' en posició intermitja d'oració (no final ni inici): per aquests segments, el límit d'inici és la pròpia marca del format del segment, que és inambigua, és a dir, no ha d'ocórrer precedida de pauses sintàctiques perquè es pugui determinar la seva funció com a límit de segment. En aquest tipus de segments, les comes tan sols es requereixen per marcar el final del segment, tot i que el límit final de segment, una unitat de tipus 'limit', també pot venir donat per un segment de tipus 'fort' o 'feble'. Com en la regla anterior, les combinacions de segments no desitjades que possibilita aquesta regla en combinació amb les regles d'oració (regles 4-7) no es dónen en els textos, ja que caldria trobar dos signes de puntuació adjacents, cosa que no es dóna[1] .

12. mig ==> feble, limit.

    Defineix el context d'ocorrença d'un segment del tipus 'feble' en posició intermitja d'oració, que és idèntic al context dels segments del tipus 'fort'. La diferència entre aquests dos tipus de segment rau en la seva fiabilitat: els segments de tipus 'fort' són inequívocament segments, mentre que els segments de tipus 'feble' són segments possibles.

13. mig ==> text, limit.

    Dóna la possibilitat que se succeeixin diversos segments de text sense format separats per unitats de tipus 'limit' en la posició central d'oració.

14. mig ==> text.

    Defineix una unitat del tipus 'text' com al contingut mínim de la posició intermitja d'oració, que, al seu torn, és l'única posició necessària per la constitució de segment. Per tant, el contingut mínim necessari perquè es constitueixi una oració (regla 7) és 'text' i la marca del límit d'oració, el punt.

15. text ==> parentetic, ws.

    Les regles 15 a 23 defineixen el contingut de la unitat 'text', que consisteix en totes les combinacions possibles de la unitat 'ws' (cadena de paraules) i els segments de tipus parentètic (és a dir, els fragments de text entre parèntesis, cometes o guions amb funció de parèntesi). Aquesta unitat és la unitat bàsica de contingut dels segments no parentètics, i el contingut mínim d'una oració.

16. text ==> literal, ws.

17. text ==> ws, parentetic, ws.

18. text ==> ws, literal, ws.

19. text ==> ws, parentetic.

20. text ==> ws, literal.

21. text ==> parentetic.

    El contingut mínim d'una oració pot ser un segment de tipus 'parentetic' (un fragment de text entre parèntesis o guionets).

22. text ==> literal.

[1] Excepte en les abreviatures, que són reconegudes com a tals per l'analitzador morfològic i, per tant, els punts no es consideren com a signes de puntuació sinó com a part de l'ítem lèxic en què es troben.

El contingut mínim d'una oració pot ser un segment de tipus 'literal' (un fragment de text entre cometes).

23. text ==> ws.

    El contingut mínim d'una oració pot ser un segment de tipus 'ws' (cadena de paraules).

24. limit ==> coma.

    Les regles 24 a 26 defineixen el contingut de la unitat 'coma', que pot ser una coma, un segment de tipus fort o un segment de tipus feble, ambdós dellimitats per una unitat de tipus 'limit', al seu torn.

25. limit ==> fort, limit.

26. limit ==> feble, limit.

27. interior ==> coma, format, coma, text.

    Les regles 27, 28 i 29 defineixen el contingut de la unitat 'interior', que dóna compte del contingut dels segments amb format que no està definit per les corresponents marques de segment amb format. La regla 27 és necessària, perquè, immediatament després d'una marca de segment, sovint es troben segments enniuats. Amb un estudi de corpus més aprofundit, tractarem d'ampliar els tipus d'unitats que poden trobar-se entre comes immediatament després de marca de segment, de manera que, amb les comprovacions necessàries, podríem afegir la regla:
    interior ==> coma, text, coma, text.

    ***exemple*:** oracio(mig(text(*La regla 27 és necessària*)), final(format(*perquè* interior(, format(*immediatament després d'una marca de segment*) , text(*sovint es troben segments enniuats*)).)

28. interior ==> text, coma, dins, coma, text.

    Dóna compte de la possibilitat que, dins d'un segment amb format, hi hagi segments menors, també amb format, enniuats. No donem la possibilitat que hi hagi tot tipus de segment enniuat dins d'un segment amb format perquè aleshores es donen anàlisis incorrectes: per les característiques de l'analitzador que utilitzem[2], queda enniuada tota la seqüència de paraules a partir d'una marca de segment amb format dins d'aquest segment amb format.

    ***exemple*:** oracio(inicial(format(*A pesar de* text(*haberlo estudiado*, dins(sadv(*desgraciadamente*)), *muy bien*,)), mig(text(*no pudo aprobar*)).)

29. interior ==> text.

    Defineix el contingut mínim d'una unitat del tipus 'interior', que ha de ser, com a mínim, 'text'.

30. fort ==> delfort, interior.

    Defineix el format dels segments amb format de tipus 'fort'. Aquests segments són aquells introduïts per marques de tipus 'delfort', que no requereixen la presència de pauses sintàctiques per la seva interpretació inequívoca com a límits de segment.

[2]L'analitzador TACAT es basa en mètodes 'greedy'. Aquests mètodes apliquen la regla més llarga o més profunda sempre que tenen la possibilitat d'aplicar dues o més regles en un moment de l'anàlisi.

***exemple*:** oracio(mig(text(*Le dieron el premio*) fort (*aunque* text(*no lo merecía*)).)

31. feble ==> delfeble, interior.

    Defineix el format dels segments amb format de tipus 'feble'. Aquests segments són aquells introduïts per marques de tipus 'delfeble', que, sense la presència de pauses sintàctiques, identifiquen segments possibles, però no segurs.

32. dins ==> prepo.

    Les regles 32 a 35 llisten els tipus de segment amb format de tipus 'dins', que poden trobar-se enniuats dins d'un segment amb format.

33. dins ==> adver.

34. dins ==> aposicio.

35. dins ==> adjec.

36. inicial ==> prepo, coma, format.

    Les regles 36 a 45 llisten els tipus de segment amb format que poden trobar-se en posició inicial d'oració: tots els segments del tipus 'format' excepte els de tipus 'rel' (relatius), 'aposicio' i 'adjec', que no poden ocórrer en aquesta posició amb l'estatus de segment. En aquesta posició, les marques de format de segment són inambiguament també marques de límit inicial de segment per la seva ocurrència immediatament després de pausa sintàctica forta, el punt de final d'oració. El límit de final de segment ve donat per la coma que separa la posició inicial de la resta de l'oració (regles 4 i 5). A més, en aquesta posició, tot segment amb format que segueixi immediatament el segment amb format d'inici d'oració tindrà inequívocament estatus de segment, ja que molt improbablement pot estar subcategoritzat pel predicat de l'oració[3].

37. inicial ==> prepo.

38. inicial ==> adver, coma, format.

39. inicial ==> adver.

40. inicial ==> impers, coma, format.

41. inicial ==> impers.

42. inicial ==> marcat, coma, format.

43. inicial ==> marcat.

44. inicial ==> coordinat, coma, format.

45. inicial ==> coordinat.

46. final ==> coma, rel.

[3] L'únic tipus de segment amb format que pot ocórrer immediatament després d'un segment inicial amb format seguit de coma i alhora estar subcategoritzat pel verb és una frase subordinada completiva. Aquest tipus de frases no es consideren segments amb format, però, sovint, la seva anàlisi morfosintàctica les fa confondre amb un segment amb format del tipus 'rel' (frase de relatiu). Aquests casos són pocs i poden ser corregits en el procés d'interpretació retòrica, però intentarem aprofundir l'estudi de les claus que ens permetin determinar quan aquests segments són subcategoritzats i quan no ho són.

Les regles 46 a 56 llisten els tipus de segment amb format que poden trobar-se en posició final d'oració: tots els segments del tipus 'format', els segments dels tipus 'fort' i 'feble' i els segments del tipus 'parenteticpunt', que tan sols ocorren a final d'oració. Aquests tres darrers tipus ocorren sense cap coma que faci inambigua una marca de segment, mentre que la resta requereixen una coma que faci inambigües les marques de format de cada tipus de segment com a marques de límit inicial de segment. La marca de límit final de segment és, en la posició final d'oració, el punt que tanca l'oració (regles 4 i 6).

47. final ==> coma, prepo.

48. final ==> coma, adver.

49. final ==> coma, impers.

50. final ==> coma, aposicio.

51. final ==> coma, adjec.

52. final ==> coma, marcat.

53. final ==> coma, coordinat.

54. final ==> parenteticpunt.

55. final ==> fort.

56. final ==> feble.

57. format ==> rel.

    Les regles 57 a 64 llisten els segments que poden estar en la posició 'format', és a dir, enniuats i delimitats per comes o marques de inambigües de límit de segment (delfort i delfeble) en posició intermitja d'oració.

58. format ==> prepo.

59. format ==> adver.

60. format ==> impers.

61. format ==> aposicio.

62. format ==> adjec.

63. format ==> marcat.

64. format ==> coordinat.

65. coordinat ==> coord, interior.

    Defineix el contingut i el format dels segments de tipus 'coordinat', que segueix el format estàndard dels segments amb format: una marca, que explicita el límit d'inici del segment i el tipus de segment de què es tracta, seguida d'una unitat de tipus 'interior'. Els límits d'inici i final de segment vénen donats per la posició en què ocorre el segment amb format: si és 'inicial' (principi d'oració, entre punt i coma, delfort o delfeble), 'final' (final d'oració, entre coma i punt) o 'format' (posició intermitja d'oració, entre coma i coma, delfort o delfeble). En els segments de tipus coordinat, la marca que determina el tipus de segment són marcadors del discurs del tipus 'conjunció'.

66. marcat ==> subordinant, interior.

    Defineix els segments amb format de tipus 'marcat'. En aquests segments, les marques de tipus de segments són marcadors del discurs del tipus 'subordinant', independentment que siguin del tipus 'delfort' o 'delfeble'.

67. impers ==> ger, interior.

    Les regles 67 a 70 defineixen els segments de tipus 'impers' (frases no personals). La marca de límit d'inici de segment és la forma no personal d'un verb, precedida de pausa sintàctica[4], i la marca final pot ser una pausa sintàctica o bé una marca del tipus 'delfort' o 'delfeble', que també ve donada per la posició en què apareixen els segments amb format. En aquests segments, la marca del tipus de segment i el contingut estan molt estretament lligats, perquè es tracta d'estructures sintàctiques, i, per tant, la marca de límit de segment constitueix una part essencial del seu contingut, i no tan sols del seu format. Aquesta regla defineix una frase no personal de gerundi, amb complements verbals.

68. impers ==> ger.

    Defineix una frase no personal de gerundi, i restringeix el contingut mínim d'un segment d'aquest tipus a la forma no personal.

69. impers ==> part, interior.

    Defineix una frase no personal de participi, amb complements verbals.

70. impers ==> part.

    Defineix una frase no personal de gerundi, i restringeix el contingut mínim d'un segment d'aquest tipus a la forma no personal.

71. rel ==> relatiu, interior.

    Com en el cas dels segments de tipus 'impers', en els segments de tipus 'rel' (frases de relatiu), la marca de límit inicial de segment i de tipus de segment, el pronom de relatiu, és, alhora, part essencial del contingut del segment, perquè es tracta d'una estructura sintàctica en què el pronom de relatiu hi té un paper central. A diferència dels segments no personals, però, no és possible de trobar un segment composat només de la marca, sinó que sempre ha d'haver-hi una unitat de tipus 'interior'. Els límits d'aquests segments es determinen pels contextos en què es donen els segments amb format.

72. prepo ==> deliprep, interior.

    Les regles 72 a 77 llisten els diferents tipus de segments que es poden formar a partir de preposicions o sintagmes preposicionals. Els límits propis dels segments amb format determinen la seva inambigüitat. Es defineixen com a segments del tipus 'prepo': els segments introduïts per una preposició del tipus 'deliprep' (amb contingut retòric explícit), els grups preposicionals i els segments caracteritzats per una preposició i un infinitiu.

73. prepo ==> grup-sp.

74. prepo ==> sp-de.

[4]La posició de pausa sintàctica ve donada per la posició en què apareixen aquest tipus de segments.

75. prepo ==> grup-sp, grup-sp.

76. prepo ==> prep, infinitiu, interior.

77. prepo ==> prep, infinitiu.

78. adjec ==> sadjec.

    Com en el cas dels segments preposicionals, en aquest segment, la marca de límit i tipus de segment i el contingut del segment són el mateix: un grup adjectiu. El context de pausa sintàctica en què es donen els segments amb format el caracteritza inequívocament com a segment.

79. adver ==> sadv.

    Com en el cas dels segments preposicionals i adjectius, en aquest segment, la marca de límit i tipus de segment i el contingut del segment són el mateix: un grup adverbial. El context de pausa sintàctica en què es donen els segments amb format el caracteritza inequívocament com a segment.

80. adver ==> adverbial.

    Especificació de la regla anterior per a les unitats de tipus 'adverbial', marcadors del discurs amb un contingut retòric explícit però amb el comportament sintàctic i segmental característic dels grups adverbials, per la qual cosa es regeix per les mateixes regles segmentals. Malgrat que no considerem aquestes unitats com a segment, les identifiquem per tal que no produeixin errors en la identificació d'altres segments.

81. aposicio ==> nompropi.

    Com en el cas dels segments preposicionals, adjectius i adverbials, en aquest segment, la marca de límit i tipus de segment i el contingut del segment són el mateix: una entitat amb nom. El context de pausa sintàctica en què es donen els segments amb format el caracteritza inequívocament com a segment.

82. subordinant ==> cs0x.

    Les regles 82, 83 i 84 creen la categoria unificadora de delimitador 'subordinant', en què no és rellevant distingir entre funció delimitadora forta o feble. Aquesta categoria redueix el nombre de regles necessàries per donar compte dels segments de tipus 'marcat', alhora que unifica el tractament de certes unitats que, en el context en què es pot donar la unitat 'subordinant', presenten comportaments idèntics.

83. subordinant ==> cs0d.

84. subordinant ==> deliprep, subord.

85. deliprep ==> fortprep.

    Les regles 85 i 86 creen la categoria unificadora de delimitador 'deliprep', en què no és rellevant distingir entre funció delimitadora forta o feble. Aquesta categoria redueix el nombre de regles necessàries per donar compte de les unitats de tipus 'prepo' i 'subordinant', alhora que unifica el tractament de certes unitats que, en el context en què es pot donar la unitat 'deliprep', presenten comportaments idèntics.

86. deliprep ==> febleprep.

87. delfort ==> cs0x.

    Les regles 87 a 90 llisten les marques de tipus 'delfort', que són marques inambigües de límit inicial de segment, és a dir, que marquen l'inici d'un segment sense necessitat que es doni el context de pausa sintàctica inicial necessari per al reconeixement inequívoc d'un segment amb format. En aquesta regla, es categoritzen les paraules amb l'etiqueta morfològica cs0x, marcadors del discurs de tipus subordinant amb funció delimitadora inambigua, com a marques 'delfort'.

88. delfort ==> fortprep, subord.

    Defineix la combinació d'una preposició de tipus 'fortprep' amb una conjunció subordinant del tipus 'subord' (que) com a marca de límit de segment del tipus 'delfort'.

89. delfort ==> fortprep, infinitiu.

    Defineix la combinació d'una preposició de tipus 'fortprep' amb un infinitiu com a marca de límit de segment del tipus 'delfort'.

90. delfort ==> fortprep.

    Defineix una preposició del tipus 'fortprep' com a marca de límit de segment del tipus 'delfort'. Per definir les preposicions no es recorre directament a l'etiqueta morfològica associada a la paraula, com en el cas de la regla 87, perquè no hem pogut associar una única etiqueta morfològica als marcadors del discurs de tipus prepositiu amb funció delimitadora inambigua. Les etiquetes morfològiques amb què treballem no tenen prou capacitat expressiva per donar compte alhora de la funció retòrica i del contingut morfològic de les formes prepositives contretes amb l'article masculí singular. Per això hem hagut de crear la categoria 'fortprep', que engloba algunes d'aquestes unitats amb una etiqueta específica (sps0x) i la resta amb regles del tipus fortprep ==> spcms (a_pesar_del).

91. delfeble ==> cs0d.

    Les regles 91 a 97 llisten les marques del tipus 'delfeble', que són marques de límit de segment possible. Com en el cas de les marques 'delfort', tampoc no es requereix un context de pausa sintàctica per al seu reconeixement inequívoc, però, a diferència de les marques 'delfort', no identifiquen segments amb un estatus consolidat, sinó tan sols probable.. En aquesta regla, es categoritzen les paraules amb l'etiqueta morfològica cs0d, marcadors del discurs de tipus subordinant amb doble funció delimitadora, com a marques 'delfeble'.

92. delfeble ==> febleprep, subord.

    Defineix la combinació d'una preposició de tipus 'febleprep' amb una conjunció subordinant del tipus 'subord' (que) com a marca de límit de segment del tipus 'delfort'.

93. delfeble ==> febleprep, infinitiu.

    Defineix la combinació d'una preposició de tipus 'febleprep' amb un infinitiu com a marca de límit de segment del tipus 'delfort'.

94. delfeble ==> febleprep.

Defineix una preposició del tipus 'febleprep' com a marca de límit de segment del tipus 'delfeble'. Com en el cas de les unitats de tipus 'fortprep' (regla 90) la categoria 'febleprep', que engloba algunes d'aquestes unitats amb una etiqueta específica (sps0d) i la resta amb regles del tipus febleprep ==> spcms (a_pesar_del).

95. delfeble ==> spcms(al), infinitiu.

    En aquesta regla, la combinació de la preposició 'al' amb un infinitiu és categoritzada com a 'delfeble'. Aquesta preposició només té funció delimitadora possible en la combinació amb infinitiu. És per això que no podem caracteritzar-la sistemàticament com a 'delfeble', però amb aquesta regla recollim el seu significat causal quan tenim prou evidència per determinar-lo automàticament (immediatament seguida d'infinitiu).

96. delfeble ==> sps00(por), infinitiu.

    Per les mateixes raons que en la regla anterior, la combinació de la preposició 'por' amb un infinitiu és categoritzada com a 'delfeble'.

97. delfeble ==> sps00(para), infinitiu.

    Per les mateixes raons que en la regla anterior, la combinació de la preposició 'para' amb un infinitiu és categoritzada com a 'delfeble'.

98. parentetic ==> Fpa((), format, migs, final, Fpt()).

    Les regles 98 a 113 donen compte dels segments de tipus parentètic, és a dir, limitats per parèntesis, guionets amb funció de parèntesis i cometes. Aquests segments han de ser tractats de manera diferent als segments amb format perquè:

    (a) tenen un contingut recursiu: poden contenir qualsevol tipus de segment, excepte oració, per la qual cosa cal tornar a enunciar les regles d'oració substituïnt la característica definitòria d'oració, el punt, per la característica definitòria dels segments parentètics, les marques de parentització. En el cas dels segments parentètics que s'inicien i acaben amb signes iguals (cometes i guionets), cal especificar també que poden contenir qualsevol tipus de segment menys un de parentètic, ja que, si no és així, les característiques de l'analitzador que fem servir[5] donarà com a enniuats els uns dins dels altres diferents segments parentètics que es trobin en la mateixa oració.
    (b) presenten una distribució totalment impredible, és a dir, poden aparèixer en qualsevol posició. Aquest tipus de distribució és idèntica a la distribució de les unitats de tipus 'w' (paraula), per la qual cosa es dóna compte de la seva distribució en les regles de la unitat 'text' (regles 15 a 23).

    Les regles 98 a 101 especifiquen el format i contingut dels segments parentètics limitats per parèntesis ('(' i ')'), que, com que comencen i acaben amb signes diferents, no contenen unitats parentètiques iguals a elles mateixes, així, per descriure'n el contingut podem recórrer a les unitats amb què descrivim el contingut de tots els segments. La única diferència és que els segments amb format que en posició inicial no estan restringits als del tipus 'inicial', sinó que poden donar-se tots els tipus de segment 'format'.

99. parentetic ==> Fpa((), format, coma, migs, Fpt()).

100. parentetic ==> Fpa((), migs, final, Fpt()).

[5] Com hem dit en una nota anterior, l'analitzador TACAT es basa en mètodes 'greedy'.

101. parentetic ==> Fpa((), migs, Fpt()).

102. parentetic ==> guionet, format, coma, migsp, final, guionet.

    Les regles 102 a 113 donen compte del contingut dels segments parentètics que utilitzen signes iguals com a límit d'inici i de final (" i -). En aquests casos, cal especificar que no poden contenir segments parentètics, per la qual cosa cal reformular les regles d'oració emprant unitats que no continguin segments parentètics, és a dir: cal substituir les unitats de tipus 'text' per seqüències de paraules sense parèntesis ('ws'). No obstant, tan sols hem realitzat aquesta reformulació en els segments de tipus 'mig', que han estat traduïts als segments de tipus 'migp' per als interiors d'aquests segments parentètics. No hem trobat que es donin estructures parentètiques dintre dels segments amb format dins d'un segment parentètic, per la qual cosa hem decidit de no crear el seu equivalent sense parèntesi.

103. parentetic ==> guionet, format, coma, migsp, guionet.

104. parentetic ==> guionet, migsp, final, guionet.

105. parentetic ==> guionet, migsp, guionet.

106. parenteticpunt ==> guionet, format, coma, migsp, final.

107. parenteticpunt ==> guionet, format, coma, migsp.

108. parenteticpunt ==> guionet, migsp, final.

109. parenteticpunt ==> guionet, migsp

110. literal ==> cometa, format, migsp, final, cometa.

111. literal ==> cometa, format, migsp, cometa.

112. literal ==> cometa, migsp, final, cometa.

113. literal ==> cometa, migsp, cometa.

114. migsp ==> migp, migsp.

    Les regles 114 a 121 presenten la reformulació del tipus de segment 'mig' perquè no s'enniuin segments parentètics dins altres segments parentètics: les unitats de tipus 'text' han estat substituïdes per 'ws' (cadena de paraules). Aquesta regla dóna recursivitat en el contingut dels segments parentètics.

115. migsp ==> migp.

    Restringeix el contingut dels segments parentètics a un mínim d'una unitat 'migp'.

116. migp ==> ws, coma, format, limit.

    Contempla la possiblitat que un segment amb format es trobi enniuat en posició intermitja, no inicial ni final, d'un segment parentètic (vegeu regla 10).

117. migp ==> fort, limit.

    Contempla la possibilitat que un segment amb format de tipus 'fort' es trobi enniuat en posició intermitja d'un segment parentètic (vegeu regla 11).

118. migp ==> feble, limit.

Contempla la possibilitat que un segment amb format de tipus 'feble' es trobi enniuat en posició intermitja d'un segment parentètic (vegeu regla 12).

119. migp ==> ws, limit.

Contempla la possibilitat que se succeeixin diverses cadenes de paraules sense format separades per una unitat de tipus 'limit' en posició intermitja d'un segment parentètic (vegeu regla 13)

120. migp ==> ws.

Restringeix el contingut d'un segment parentètic a un mínim de 'ws' (cadena de paraules).

121. ws ==> w, ws.

Defineix la recursivitat de la unitat 'ws' (cadena de paraules), de manera que doni compte d'una seqüència d'un nombre indeterminat de paraules (w). Són unitats del tipus 'w' totes les unitats lèxiques, grups sintàctics i marques de delimitació de segments ambigües, i no ho són els signes de puntuació ni les marques de delimitació de segments inambigües.

122. ws ==> w.

Restringeix el contingut de la unitat 'ws' (cadena de paraules), que ha de contenir, com a mínim, una unitat 'w' (paraula).

Apèndix B

Annex II: Mostra de text segmentat

Tot seguit adjuntem una mostra de text, en la seva sortida del segmentador automàtic.

En aquest exemple podem veure com els segments coordinats no es solucionen de manera adequada. En el primer cas, la conjunció 'ni' està coordinant els dos segments '*un violador nunca será mister, maestro o profesor*' i '*un secuestrador, torturador o ladrón recibirá el trato de cortesía que supone anteponer a su nombre la voz don*'. La coma que tanca el primer segment dóna el context necessari perquè la conjunció 'ni' sigui interpretada com a delimitador inambigu. Però aquest segment es tanca de manera prematura, per una mala interpretació d'una de les comes de la llista que conté. Aquesta llista té nivell lèxic, i per tant, no té una funció discursiva, sinó oracional. El segon cas és paral·lel al primer. Com queda palès, la solució de la coordinació és una de les qüestions més difícils de resoldre en el PLN, perquè resulta molt complex de sistematitzar de manera satisfactòria la gran varietat d'elements que pot unir.

També trobem en aquesta mostra un exemple de segment de tipus 'feble' amb funció oracional. Per la seva ambigüitat, aquest tipus de segments s'identifiquen però no es consideren segments discursivament significatius.

```
Una vieja norma periodística, que no sé si ha sido abolida de los manuales,
dice que los criminales han de ser tratados sin consideración social o de
rango alguna. Así, por ejemplo,un asesino convicto y confeso jamás será
tratado de señor;un violador nunca será mister, maestro o profesor, ni un
secuestrador, torturador o ladrón recibirá el trato de cortesía que supone
anteponer a su nombre la voz don, a menos que sea una ironía o se quiera
dar la idea de que hablamos de un mafioso. ¿Es Pinochet un criminal? ¡Sí!,
qué duda cabe. Y también complotador,golpista, traidor, perjuro. ¿Cómo
puede, pues, seguírsele tratando de general, cuando en rigor protocolario
el grado reemplaza al adjetivo señor y además está jubilado y sujeto a proceso
de extradición?
```

paragraf_[oracio_[**Una**_pi3fs000 **vieja**_aq0fs00 **norma**_ncfs000 **periodística**_aq0fs00 **,**_Fc rel_[**que**_pr3cn000 **no**_rg000 **sé**_vmip1s0] feble_[delfeble_[**si**_cs0d] **ha**_vaip3s0 **sido**_vap00nc **abolida**_vmp00sf **de**_sps00 **los**_tdmp0 **manuales**_ncmp000] **,**_Fc **dice**_vmip3s0 **que**_cs00 **los**_tdmp0 **criminales**_nccp000 **han**_vaip3p0 **de**_sps00 **ser**_van0000 **tratados**_vmp00pm **sin**_sps00 **consideración**_ncfs000 **social**_aq0cs00 **o**_cc00 **de**_sps00 **rango**_ncms000 **alguna**_pi3fs000 **.**_Fp] oracio_[adver_[adverbial_[**Así**_rg00d]] **,**_Fc adver_[adverbial_[**por_ejemplo**_rg00d]] **,**_Fc **un**_di3ms00 **asesino**_ncms000 **convicto**_aq0ms00 **y**_cc00 **confeso**_aq0ms00 **jamás**_rg000 **será**_vaif3s0 **tratado**_vmp00sm **de**_sps00 **señor**_ncms000 **;**_Fx] oracio_[**un**_mcms0d **violador**_ncms000 **nunca**_rg000 **será**_vaif3s0 **mister**_vmp0000 **,**_Fc **maestro**_ncms000

o_cc00 **profesor**_ncms000 **,**_Fc coordinat_[conjuncio_[**ni**_cc00] **un**_mcms0d **secuestrador**_ncms000] ,_Fc] **torturador**_aq0ms00 **o**_cc00 **ladrón**_ncms000 **recibirá**_vmif3s0 **el**_tdms0 **trato**_ncms000 **de**_sps00 **cortesía**_ncfs000 **que**_pr3cn000 **supone**_vmip3s0 **anteponer**_vmn0000 **a**_sps00 **su**_dp3cs00 **nombre**_ncms000 **la**_tdfs0 **voz**_ncfs000 **don**_ncms000 ,_Fc fort_[delfort_[**a**_**menos**_**que**_cs0x] **sea**_vasp1s0 **una**_mcfs0d **ironía**_ncfs000 **o**_cc00 **se**_pp3cn000 **quiera**_vmsp1s0 **dar**_vmn0000 **la**_tdfs0 **idea**_ncfs000 **de**_sps00 **que**_pr3cn000 **hablamos**_vmip1p0 **de**_sps00 **un**_di3ms00 **mafioso**_ncms000] **.**_Fp] oracio_[¿_Fia **Es**_vaip3s0 **Pinochet**_np00sp0 **un**_z **criminal**_nccs000 **?**_Fit] oracio_[**¡**_Faa **Sí**_rg000 **!**_Fat] limit_[**,**_Fc] oracio_[**qué**_dt3cn00 **duda**_ncfs000 **cabe**_vmip3s0 **.**_Fp] oracio_[coordinat_[conjuncio_[**Y**_cc00 **también**_rg000 **complotador**_ncms000] ,_Fc **golpista**_aq0cs00 ,_Fc **traidor**_aq0ms00 **,**_Fc **perjuro**_aq0ms00 **.**_Fp] oracio_[**¿**_Fia **Cómo**_np00sp0 **puede**_vmip3s0 **,**_Fc adver_[adverbial_[**pues**_rg00d]] **,**_Fc **seguírsele**_vmn0000 **tratando**_vmg0000 **de**_sps00 **general**_aq0cs00 **,**_Fc marcat_[subordinant_[**cuando**_cs0d] adverbial_[**en**_**rigor**_rg00d] **protocolario**_aq0ms00 **el**_tdms0 **grado**_ncms000 **reemplaza**_vmip3s0 **al**_spcms **adjetivo**_ncms000 **señor**_aq0ms00] fort_[delfort_[**y**_**además**_cs0x] **está**_vmip3s0 **jubilado**_vmp00sm **y**_cc00 **sujeto**_aq0ms00 **a**_sps00 **proceso**_ncms000 **de**_sps00 **extradición**_ncfs000] **?**_Fit]]

.

ORACIO_[
Una_pi3fs000
vieja_aq0fs00
norma_ncfs000
periodística_aq0fs00
,_Fc
REL_[
que_pr3cn000
no_rg000
sé_vmip1s0
]
FEBLE_[
DELFEBLE_[
si_cs0d
]
ha_vaip3s0
sido_vap00nc
abolida_vmp00sf
de_sps00
los_tdmp0
manuales_ncmp000
]
,_Fc
dice_vmip3s0
que_cs00
los_tdmp0
criminales_nccp000
han_vaip3p0
de_sps00
ser_van0000
tratados_vmp00pm
sin_sps00
consideración_ncfs000
social_aq0cs00

o_cc00
de_sps00
rango_ncms000
alguna_pi3fs000
._Fp
]
ORACIO_[
ADVER_[
ADVERBIAL_[
Así_rg00d
]
]
,_Fc
ADVER_[
ADVERBIAL_[
por_ejemplo_rg00d
]
]
,_Fc
un_di3ms00
asesino_ncms000
convicto_aq0ms00
y_cc00
confeso_aq0ms00
jamás_rg000
será_vaif3s0
tratado_vmp00sm
de_sps00
señor_ncms000
;_Fx
]
ORACIO_[
un_mcms0d
violador_ncms000
nunca_rg000
será_vaif3s0
mister_vmp0000
,_Fc
maestro_ncms000
o_cc00
profesor_ncms000
,_Fc
COORDINAT_[
CONJUNCIÓ_[
ni_cc00
]
un_mcms0d
secuestrador_ncms000
]
,_Fc
torturador_aq0ms00
o_cc00
ladrón_ncms000
recibirá_vmif3s0

el_tdms0
trato_ncms000
de_sps00
cortesía_ncfs000
que_pr3cn000
supone_vmip3s0
anteponer_vmn0000
a_sps00
su_dp3cs00
nombre_ncms000
la_tdfs0
voz_ncfs000
don_ncms000
,_Fc
FORT_[
DELFORT_[
a_menos_que_cs0x
]
sea_vasp1s0
una_mcfs0d
ironía_ncfs000
o_cc00
se_pp3cn000
quiera_vmsp1s0
dar_vmn0000
la_tdfs0
idea_ncfs000
de_sps00
que_pr3cn000
hablamos_vmip1p0
de_sps00
un_di3ms00
mafioso_ncms000
]
._Fp
]
ORACIO_[
¿_Fia
Es_Pinochet_np00sp0
un_z
criminal_nccs000
?_Fit
]
ORACIO_[
¡_Faa
Sí_rg000
!_Fat
]
LIMIT_[
,_Fc
]
ORACIO_[
qué_dt3cn00
duda_ncfs000

cabe_vmip3s0
._Fp
]
ORACIO_[
COORDINAT_[
CONJUNCIO_[
Y_cc00
]
también_rg000
complotador_ncms000
]
,_Fc
golpista_aq0cs00
,_Fc
traidor_aq0ms00
,_Fc
perjuro_aq0ms00
]
._Fp
]
ORACIO_[
¿_Fia
Cómo_np00sp0
puede_vmip3s0
,_Fc
ADVER_[
ADVERBIAL_[
pues_rg00d
]
]
,_Fc
seguírsele_vmn0000
tratando_vmg0000
de_sps00
general_aq0cs00
,_Fc
MARCAT_[
SUBORDINANT_[
cuando_cs0d
]
ADVERBIAL_[
en_rigor_rg00d
]
protocolario_aq0ms00
el_tdms0
grado_ncms000
reemplaza_vmip3s0
al_spcms
adjetivo_ncms000
señor_aq0ms00
]
FORT_[
DELFORT_[
y_además_cs0x

]
está_vmip3s0
jubilado_vmp00sm
y_cc00
sujeto_aq0ms00
a_sps00
proceso_ncms000
de_sps00
extradición_ncfs000
]
]
?_Fit
]

Apèndix C

Annex II: Instruccions per Segmentació Humana

Tot seguit adjuntem les instruccions que es van donar als informants per a l'experiment de segmentació humana:

Es tracta de dividir el text en els trossos que tu consideris que són una **unitat mínima de signficat**, és a dir, que té un significat independent, que podria tallar-se de la resta del text per treure'l del text o per posar-lo a part.

Us dono poques indicacions perquè necessito saber què penseu vosaltres, no pas com enteneu vosaltres el que jo penso. Però si voleu, podeu tenir en compte que:

– la divisió del text s'ha d'utilitzar per resumir aquest text més endavant, o sigui que una unitat mínima de significat s'hauria de poder utilitzar, amb poques adaptacions, per posar-la en un text.

  exemple:

  No Útil: *La decisión significa* ***[que el procesado no deberá posar para ser fotografiado de frente y perfil]*** *, ni estampar sus huellas dactilares en una ficha policial, como debe hacerlo en Chile todo ciudadano sometido a un proceso.*

  Útil: ***[*** *La decisión significa que el procesado no deberá posar para ser fotografiado de frente y perfil* ***]***, ***[*** *ni estampar sus huellas dactilares en una ficha policial* ***]***, ***[*** *como debe hacerlo en Chile todo ciudadano sometido a un proceso* ***]*** . ***[*** *En cambio, las fotos y huellas de Pinochet,* ***[*** *archivadas en el Servicio de Registro Civil e Identificación* ***]*** *, se adjuntarán al expediente del proceso* ***]***, ***[*** *sin necesidad de que el procesado comparezca* ***]***.

– els signes de puntuació, les oracions subordinades, les paraules com pero, es decir, etc., poden ser una mica difícils perquè no sabeu si posar-les amb un segment de text, amb un altre, o soles. No us hi encaparreu, feu-ho com millor us sembli. La vostra intuïció és important!

– pot ser que trobeu segments dins d'altres segments, com per exemple:

  El ex dictador chileno Augusto Pinochet*, procesado por su presunta responsabilidad en violaciones a los derechos humanos,* ***será fichado 'en ausencia',*** *es decir, sin hacerle fotos ni tomarle huella.*

  Aquests casos, podeu marcar-los així:

> *El ex dictador chileno Augusto Pinochet* ***[****, procesado por su presunta responsabilidad en violaciones a los derechos humanos,* ***]*** *será fichado 'en ausencia', # es decir, sin hacerle fotos ni tomarle huella.*

I podeu posar tants nivells com volgueu, com per exemple:

> *El ex dictador chileno Augusto Pinochet* ***[****, procesado****[****, después de su estancia en Londres,****]*** *por su presunta responsabilidad en violaciones a los derechos humanos,* ***]*** *será fichado 'en ausencia', # es decir, sin hacerle fotos ni tomarle huella.*

Però no cal que us compliqueu gaire!! També seria útil:

> *El ex dictador chileno Augusto Pinochet* ***[****, procesado, después de su estancia en Londres, por su presunta responsabilidad en violaciones a los derechos humanos,* ***]*** *será fichado 'en ausencia', # es decir, sin hacerle fotos ni tomarle huella.*

Apèndix D

Annex IV: Marcadors del Discurs

Presentem adjunta una mostra del lèxic de marcadors del discurs que hem desenvolupat. La base de dades amb què treballem conté 610 formes, però aquí tan sols presentem les 388 bàsiques. D'una banda, no incloem les extensions de forma dels marcadors que segueixen patrons regulars. Per exemple, incloem '*a base de*' però no '*a base del*' ni '*a base de que*', ja que podem determinar la forma i els atributs associats a aquests dos darrers marcadors mitjançant regles de composició a partir del primer. D'altra banda, tan sols mostrem aquells marcadors amb una freqüència en corpus major a 5, la resta són marcadors molt poc utilitzats, que han quedat fora d'ús o que pertanyen a un registre molt culte de la llengua. En general, la descripció d'aquests marcadors és extrapolable de la d'algun dels marcadors que presentem aquí.

Mostrem en aquesta primera versió de la base de dades de marcadors del discurs la informació que s'utilitza de manera efectiva per a segmentació. Per això, no incloem tots els paràmetres que hem descrit en l'apartat 3.2.3, sinó els següents:

- **forma** del marcador: l'accés del segmentador al coneixement associat als marcadors es fa a través de la seva forma
- funció **relacional**: en aquest paràmetre tan sols determinem la funció delimitadora del marcador. Si en té, distingim entre funció discursiva inambigua (**fort**) o doble funció (**feble**), si no en té, no apliquem el valor (**nap**)
- tipus **sintàctic**: hem distingit **preposicional**, **subordinant**, **conjunció**, **adverbial**, i **no personal**. En la gramàtica de segmentació, cadascun d'aquests tipus té un comportament diferent
- comportament **segmental**: aquesta informació no és rellevant per segmentació, però l'hem especificada perquè creiem que serà útil per interpretació. Hem distingit **dreta**, **esquerra**, **ambigu**, **bidireccional** i **inclusió**.
- tipus **retòric**: incloem aquesta informació per les mateixes raons que en el cas anterior. Hem distingit **satel·litzador**, **nucleitzador**, **connector**, **organitzador** (entre els organitzadors hem distingit també **introductors**, **distribuidors** i **seqüenciadors**)
- **contingut** retòric proposat: com en els dos casos anteriors, creiem que aquesta informació pot ser útil per interpretació. No obstant, no hem aprofundit en les relacions retòriques amb contingut, i, per tant, hem donat compte del contingut retòric de manera intuïtiva, no sistemàtica. Entre els significats que hem distingit, podem observar els següents grups:
  - **circumstancial** (16): **temps** (18), **manera** (32),
  - **seqüència** (12): en aquest grup podem distingir
    * **causalitat** (3): **causa** (19), **resultat** (1), **finalitat** (15), **conseqüència** (31)

tipus sintàctic	comportament segmental	funció relacional	nombre d'ocurrències
preposicional	dreta	fort	48
subordinant	dreta	feble	44
preposicional	dreta	feble	39
subordinant	dreta	fort	35
anafòric	esquerra	-	28
adverbial	esquerra	-	20
adverbial	inclusió	-	17

Taula D.1: Combinacions més freqüents de paràmetres

    * **condicionalitat** (27), restricció (1)
    * **capacitació** (13)
- **argumentatiu**: podem distingir:
    * **reforç** argumentatiu (19)
    * contraargument: **adversació** (4), **concessió** (35)
- **comparació** (18): **paral·lelisme** (4), **contrast** (11)
- **concreció** (5), **exemple** (3)
- **resum** (11), **conclusió** (1)
- gestió de la **informació** (31): **reformulador** (22), **ordre** (23), **origen** del missatge (3)
- continguts **difícils de classificar**: **instrument** (1) (causa-manera), **excepció** (5) (contrast-condicional)
- continguts **híbrids**: conseqüència-manera, seqüència-temps, circumstància-temps

Aquests paràmetres presenten uns patrons de combinació força regulars (vegeu taula C.1).