Coreferència:
Teoria, anotació, resolució i avaluació

Marta Recasens Potau

Coreferència:
Teoria, anotació, resolució i avaluació

per

Marta Recasens Potau

Memòria presentada dins del
Programa de Doctorat *Lingüística i Comunicació*,
Bienni 2006–2008,
Departament de Lingüística General,
Universitat de Barcelona,
per optar al grau de **Doctor**

sota la direcció de

Dra. M. Antònia Martí Antonín
Universitat de Barcelona

Dr. Eduard Hovy
ISI – University of Southern California

Universitat de Barcelona
Setembre 2010

Comptar paraules no és molt revelador si no les escoltes també.

– Geoffrey Nunberg

"The I's Don't Have It", Fresh Air (November 17, 2009)

Els factors d'ús revelen que la llengua és un instrument social, orgànic i natural, no un instrument lògic i abstracte.

– Joan Bybee

Language, Usage and Cognition (2010:193)

Als meus estimats, Mercè, Eduard, Elm i Mark, per ser-hi.

Resum

Les relacions de coreferència, segons la definició més comuna, s'estableixen entre expressions lingüístiques que es refereixen a una mateixa persona, objecte o esdeveniment. Resoldre-les és una part integral de la comprensió del discurs ja que permet als usuaris de la llengua connectar les parts del discurs que contenen informació sobre una mateixa entitat. En conseqüència, la resolució de la coreferència ha estat un focus d'atenció destacat del processament del llenguatge natural, on té una tasca pròpia. Tanmateix, malgrat la gran quantitat de recerca existent, els resultats dels sistemes actuals de resolució de la coreferència no han assolit un nivell satisfactori.

La tesi es divideix en dos grans blocs. En el primer, examino tres aspectes diferents però estretament relacionats de la tasca de resolució de la coreferència: (i) l'anotació de relacions de coreferència en grans corpus electrònics, (ii) el desenvolupament de sistemes de resolució de la coreferència basats en aprenentatge automàtic i (iii) la qualificació i avaluació dels sistemes de coreferència. En el transcurs d'aquesta investigació, es fa evident que la tasca de coreferència presenta una sèrie de problemes de base que constitueixen veritables obstacles per a la seva correcta resolució. Per això, la meva aportació principal és una anàlisi crítica i alhora constructiva de diferents aspectes de la tasca de coreferència que finalment condueix, en el segon bloc de la tesi, al replantejament del concepte mateix de *coreferència*.

En primer lloc, l'anotació amb coreferència dels corpus AnCora del castellà i el català (un total de 800.000 paraules) posa al descobert, d'una banda, que el concepte de *referencialitat* no està clarament delimitat i, d'una altra, que algunes relacions observades en dades d'ús real no encaixen dins la visió de la coreferència entesa en termes dicotòmics. Tant els graus de referencialitat com les relacions que no són ni coreferencials ni no coreferencials (o que accepten totes dues interpretacions) són una de les raons principals que dificulten assolir un alt grau d'acord entre els anotadors d'aquesta tasca.

En segon lloc, els experiments realitzats sobre la contribució de més de quaranta-

cinc trets d'aprenentage automàtic a la resolució de la coreferència mostren que, tot i que el ventall de trets motivats lingüísticament porta a una millora significativa general, aquesta és més petita que l'esperada. En canvi, el senzill tret de mateix-nucli (*head match*) aconsegueix tot sol resultats prou satisfactoris. D'això se'n desprèn que es tracta d'un dels pocs trets suficientment representats per al bon funcionament de l'aprenentatge automàtic. La interacció complexa que es dóna entre els diversos factors així com el fet que el coneixement pragmàtic i del món no es deixa representar sistemàticament en forma de trets d'aprenentatge de parells de mencions són indicadors que la manera en què actualment s'aplica l'aprenentatge automàtic pot no ser especialment idònia per a la tasca de coreferència. Per això, considero que el millor model per adreçar el problema de la coreferència correspon als sistemes basats en entitats com CISTELL, que presento a la tesi. Aquest sistema permet no només emmagatzemar informació de "dins" del text sinó també recollir coneixement general i del món de "fora" del text.

En tercer lloc, altres experiments així com la tasca compartida del SemEval demostren l'existència de diversos factors que qüestionen la manera en què actualment s'avaluen els sistemes de resolució de la coreferència. Es tracta de variacions en la definició de la tasca, l'extracció de mencions a partir de l'estàndard de referència o predites automàticament, i el desacord entre els rànquings de sistemes donats per les mètriques d'avaluació més utilitzades (MUC, B^3, CEAF). La desigualtat entre el nombre d'entitats unàries i el nombre d'entitats de múltiples mencions explica el biaix de les mesures o bé cap a un dèficit o bé cap a un excés de *clusters*. La mesura BLANC que proposo, una implementació modificada de l'índex de Rand, corregeix aquest desequilibri dividint la puntuació final entre relacions de coreferència i de no coreferència.

Finalment, la segona part de la tesi arriba a la conclusió que l'abandó de la visió tradicional i dicotòmica de la coreferència és el primer pas per anar més enllà de l'estat de l'art. Amb aquest objectiu s'introdueix la noció de *quasi-identitat* i s'ubica en un model de la coreferència entesa com a *contínuum*. Des d'una perspectiva cognitiva, dono raons a favor del nivell variable de granularitat en què concebem les entitats discursives. Es postulen tres operacions de categorització –l'especificació, el reenfocament i la neutralització– que regeixen els canvis que les entitats discursives experimenten a mesura que avança el discurs i, per tant, permeten explicar les relacions de (quasi-)coreferència. Aquest nou model proporciona fonaments teòrics sòlids al problema de la coreferència tant en el camp lingüístic com en el computacional.

Abstract

Coreference relations, as commonly defined, occur between linguistic expressions that refer to the same person, object or event. Resolving them is an integral part of discourse comprehension by allowing language users to connect the pieces of discourse information concerning the same entity. Consequently, coreference resolution has become a major focus of attention in natural language processing as its own task. Despite the wealth of existing research, current performance of coreference resolution systems has not reached a satisfactory level.

The thesis is broadly divided into two parts. In the first part, I examine three separate but closely related aspects of the coreference resolution task, namely (i) the encoding of coreference relations in large electronic corpora, (ii) the development of learning-based coreference resolution systems, and (iii) the scoring and evaluation of coreference systems. Throughout this research, insight is gained into foundational problems in the coreference resolution task that pose obstacles to its feasibility. Hence, my main contribution resides in a critical but constructive analysis of various aspects of the coreference task that, in the second part of the thesis, leads to rethink the concept of *coreference* itself.

First, the annotation of the Spanish and Catalan AnCora corpora (totaling nearly 800k words) with coreference information reveals that the concept of *referentiality* is not a clear-cut one, and that some relations encountered in real data do not fit the prevailing either-or view of coreference. Degrees of referentiality as well as relations that do not fall neatly into either coreference or non-coreference—or that accept both interpretations—are a major reason for the lack of inter-coder agreement in coreference annotation.

Second, experiments on the contribution of over forty-five learning features to coreference resolution show that, although the extended set of linguistically motivated features results in an overall significant improvement, this is smaller than

expected. In contrast, the simple head-match feature alone succeeds in obtaining a quite satisfactory score. It emerges that head match is one of the few features sufficiently represented for machine learning to work. The complex interplay between factors, and the fact that pragmatics and world knowledge do not lend themselves to be captured systematically in the form of pairwise learning features, are indicators that the way machine learning is currently applied may not be well suited to the coreference task. I advocate for entity-based systems like the one presented in this thesis, CISTELL, as the model best suited to address the coreference problem. CISTELL allows not only the accumulation and carrying of information from "inside" the text, but also the storing of background and world knowledge from "outside" the text.

Third, further experiments, as well as the SemEval shared task, demonstrate that the current evaluation of coreference resolution systems is obscured by a number of factors including variations in the task definition, the use of gold-standard or automatically predicted mention boundaries, and the disagreement between the system rankings produced by the widely-used evaluation metrics (MUC, B^3, CEAF). The imbalance between the number of singletons and multi-mention entities in the data accounts for measurement biases toward either over- or under-clustering. The BLANC measure that I propose, which is a modified implementation of the Rand index, addresses this imbalance by dividing the score into coreference and non-coreference links.

Finally, the second part of the thesis concludes that abandoning the traditional categorical understanding of coreference is the first step to further the state of the art. To this end, the notion of *near-identity* is introduced within a *continuum* model of coreference. From a cognitive perspective, I argue for the variable granularity level at which discourse entities can be conceived. It is posited that three different categorization operations—specification, refocusing and neutralization—govern the shifts that discourse entities undergo as a discourse evolves and so account for (near-)coreference relations. This new continuum model provides sound theoretical foundations to the coreference problem, both for the linguistic and computational fields.

Agraïments

Per cuinar la meva tesi, he tingut l'oportunitat de treballar sota la supervisió de dos xefs de gran talent i entusiasme, M. Antònia Martí i Ed Hovy. Els estic immensament agraïda per haver-me revelat els ingredients secrets que fan que la recerca sigui fascinant i per estar sempre presents –físicament o electrònica– per ensenyar-me valuoses receptes que han passat a formar part del meu estil de cuinar. En els últims quatre anys, m'han guiat a través de tots els passos necessaris per preparar una tesi apetitosa: planificar, prendre decisions, crear, ser elegant i pacient, coure de forma lenta i delicada i servir amb estil. Gràcies per la vostra dedicació, per ajudar-me a dominar aquest art i per ser molt més que directors.

Moltes gràcies també als membres del tribunal, Costanza Navarretta, Massimo Poesio i Mariona Taulé, per haver acceptat ser al comitè avaluador. El meu sincer agraïment pels vostres comentaris i les idees que heu contribuït en diversos moments durant el procés de cocció que han ajudat a fer d'aquesta tesi una experiència nutritiva.

He estat molt afortunada de conèixer i interactuar amb altres xefs extraordinàriament hàbils que m'han suggerit receptes excel·lents per provar o que m'han donat bons consells sobre com millorar un plat. Estic en deute amb Jerry Hobbs, Horacio Rodríguez, Mihai Surdeanu, Lluís Màrquez, Ruslan Mitkov, Vincent Ng, Véronique Hoste, Antal van den Bosch, Olga Uryupina i Manu Bertran. Un enorme agraïment cap a Edgar Gonzàlez, qui sempre ha estat disposat a ajudar-me a *debugar* codi o quan m'encallava en algun problema informàtic. Gràcies per ensenyar-me a degustar Java.

També em sento molt afortunada d'haver estat envoltada de companys fantàstics en dues de les millors cuines, el Departament de Lingüística de la Universitat de

Barcelona i l'Information Sciences Institute de la University of Southern California. Un emotiu agraïment per a les meves companyes de despatx Marta Vila i Aina Peris amb qui he compartit incomptables hores d'amistat, feina, estrès, diversió i llàgrimes. He gaudit molt de la seva companyia i m'agradaria donar les gràcies a tots per crear un ambient intel·lectualment estimulant a: José Luis Ambite, Erika Barragan-Nunez, Rahul Bhagat, Oriol Borrega, Gully Burns, Congxing Cai, William Chang, Glòria de Valdivia, Steve Deneefe, Paramveer Dhillon, Victoria Fossum, Andrew Goodney, Paul Groth, Ulf Hermjakob, Dirk Hovy, Liang Huang, Tommy Ingulfsen, Zori Kozareva, Adam Lammert, Jon May, Rutu Mulkar-Mehta, Oana Nicolov, Montserrat Nofre, Anselmo Peñas, David Pynadath, Sujith Ravi, Santi Reig, John Roberto, Tom Russ, Emili Sapena, Ashish Vaswani i Rita Zaragoza. També desitjo expressar el meu sincer agraïment als meus "amics de congrés" Constantin Orasan, Laura Hasler i Marta Ruiz Costa-Jussà, amb qui he mantingut profitoses discussions tot gustant la cuina búlgara, txeca, marroquina i índia. Vull destacar a Constantin per prendre's el temps de llegir la meva tesi i aportar útils suggeriments.

M'agradaria donar les gràcies als meus estimats pares, dels qui he après a assaborir els petits moments de la vida, per ser els meus primers mestres i per animarme sempre a perseguir els meus interessos intel·lectuals. Gràcies també al meu germà Elm, qui sap realment cuinar, per fer-me l'arròs negre més deliciós que m'ha donat energia per arribar fins al final. Vull agraïr als meus grans amics no perdre el costum de reunir-nos per sopar i fer un esforç per entendre en què em passava les hores treballant. Un agraïment especial per a la Laura, la Joana, la Sara, la Lali, la Cristina, el Marc, la Carlota, la Marta, la Maria, la Laia, el Manel, l'Antonio, la Tina, i el Martí (per tots aquells gots d'*orxata*!). Finalment, però certament no menys important, gràcies Mark per aportar aquest toc màgic de felicitat i amor i per introduir el teu sabor a la meva vida.

⋆ ⋆ ⋆

Aquest treball ha estat finançat per una beca FPU (AP2006-00994) del Ministerio de Educación y Ciencia.

Índex

Índex de taules

Índex de figures

CAPÍTOL 1

Introducció

Aquesta tesi tracta de la coreferència. És la història d'un projecte que va començar amb el propòsit d'anotar un corpus amb relacions de coreferència per entrenar el primer sistema de resolució de la coreferència basat en aprenentatge automàtic per al català i el castellà, però que ha acabat desenvolupant una concepció teòrica alternativa de la coreferència. Aquesta concepció va néixer de la necessitat palesa de replantejar la definició de coreferència, de reconsiderar allò que pot ser resolt automàticament i de repensar els criteris d'avaluació. El com i el perquè d'aquest gir s'expliquen en les 200 pàgines que segueixen.

L'estructura d'aquesta tesi manté intencionadament l'ordre cronològic per tal de captar l'evolució de quatre anys que ha donat forma a les idees que presento. La idea original va anar modificant-se gradualment fins a la compleció de la present memòria. Per aquesta raó, començo situant el lector en les mateixes condicions en què jo vaig començar i així pugui seguir la progressió lògica de principi a fi.

1.1 Punt de partença

El punt de partença d'aquesta tesi va ser el problema de la resolució de la coreferència, un dels reptes del Processament del Llenguatge Natural (PLN). Es tracta d'un tema que s'acostuma a definir com "el problema d'identificar quins sintagmes nominals (SNs) o mencions del text es refereixen a la mateixa entitat del món real" (NG, 2009; STOYANOV *et al.*, 2009; FINKEL i MANNING, 2008) o bé, en termes lleugerament diferents, com "la tasca d'agrupar totes les mencions d'entitats d'un text en classes d'equivalència de manera que totes les mencions d'una classe donada es refereixin a la mateixa entitat discursiva" (BENGTSON i ROTH, 2008; DENIS i BALDRIDGE, 2009). En conseqüència, les mencions 1, 2 i 3 a (1) corefereixen, ja

que totes tres es refereixen a Eyjafjallajökull.[1]

(1) [El volcà Eyjafjallajökull, un dels més grans d'Islàndia,]$_1$ havia estat en repòs durant gairebé dos segles fins que va cobrar vida el capvespre del dia 20 de març de 2010, perceptible al principi per l'emergència d'un núvol vermell que brillava per sobre de la vasta glacera que [el]$_2$ cobreix. Durant els dies que van seguir, van brollar deus de foc d'una dotzena d'orificis d[el volcà]$_3$, que arribaven fins als 100 metres.[2]

Cal destacar que s'utilitzen diferents expressions lingüístiques –un nom propi, un pronom i un SN definit. Això, però, no és cap requisit de la coreferència: les mencions 1, 2 i 3 a (2) també serien coreferents amb l'única diferència que es perdria cohesió discursiva.

(2) [Eyjafjallajökull]$_1$ havia estat en repòs durant gairebé dos segles fins que va cobrar vida el capvespre del dia 20 de març de 2010, perceptible al principi per l'emergència d'un núvol vermell que brillava per sobre de la vasta glacera que cobreix [Eyjafjallajökull]$_2$. Durant els dies que van seguir, van brollar deus de foc d'una dotzena d'orificis d'[Eyjafjallajökull]$_3$, que arribaven fins als 100 metres.

La coreferència, com la identitat, es defineix com una relació dicotòmica: dues mencions són o bé coreferents (tenen el mateix referent) o bé no coreferents (tenen referents diferents). Els primers estudis sobre la resolució de la coreferència van derivar de la tasca germana de resolució de l'anàfora (MITKOV, 2002), que consisteix en resoldre la referència de pronoms (anafòrics) i de SNs definits la interpretació dels quals depèn d'una expressió anterior. És a dir, consisteix en identificar el seu antecedent en el text. Tot i que estan relacionades, la resolució de la coreferència fa un pas més enllà ja que requereix resoldre la referència de cadascuna de les mencions d'un text (pronoms, noms propis, SNs definits i indefinits, etc.), incloses aquelles que no depenen d'una altra expressió per a la seva interpretació.

Com a parlants de la llengua, podem trobar ràpidament i inconscient la referència de cada expressió lingüística i lligar la informació donada per aquelles expressions que es refereixen a la mateixa entitat. Tanmateix, el procés subjacent de com això té lloc encara no és clar. Explicitar de manera sistemàtica el coneixement que hi ha darrera d'aquesta pràctica és lluny de ser una qüestió trivial i d'aquí el repte que la resolució de la coreferència representa per al PLN. No obstant, hi ha un fort interès en la identificació automàtica de les relacions de coreferència ja que són clau per a "comprendre" un text i per tant esdevenen un requisit per a aplicacions del PLN com són l'extracció d'informació (MCCARTHY i LEHNERT, 1995), el resum automàtic (AZZAM *et al.*, 1999; STEINBERGER *et al.*, 2007), la cerca de respostes (MORTON, 2000; VICEDO i FERRÁNDEZ, 2006) i la traducció

[1] Com que la coreferència és un fenomen discursiu, és normalment necessari incloure un context suficientment llarg en els exemples que apareixen al llarg del capítol.

[2] The New York Times (20 d'abril de 2010).

automàtica, on cal identificar l'antecedent d'un pronom abans de traduir-lo. Les relacions de coreferència també són útils per a altres tasques com l'anàlisi de sentiments (NICOLOV *et al.*, 2008), la inferència textual (MIRKIN *et al.*, 2010; ABAD *et al.*, 2010), l'agrupament de cites bibliogràfiques i bases de dades (WICK *et al.*, 2009), la lectura automàtica (POON *et al.*, 2010), per aprendre esquemes narratius (CHAMBERS i JURAFSKY, 2008) i per recuperar arguments implícits (GERBER i CHAI, 2010; RUPPENHOFER *et al.*, 2010).

Aquesta tesi, que s'afegeix a la recerca existent en resolució de la coreferència, va néixer de l'interès per millorar els sistemes de coreferència mitjançant l'ús massiu d'aprenentatge automàtic conjuntament amb coneixement lingüístic. El meu objectiu era descobrir patrons subjacents de les relacions de coreferència i generalitzar la manera en què diferents tipus d'informació lingüística interactuen i són sospesats. A més, com que la major part de la investigació en PLN se centra en l'anglès, una segona motivació era la necessitat de desenvolupar recursos lingüístics per al català i castellà com ara un corpus anotat amb informació sobre la coreferència i un sistema de resolució de la coreferència.

Veia l'anotació de corpus com una oportunitat per introduir noves perspectives en la resolució de la coreferència. La meva hipòtesi de treball era que anotar relacions de coreferència posant èmfasi tant en la quantitat de les dades anotades com en la qualitat de l'anotació tindria un efecte positiu immediat en el model après pels mètodes d'aprenentatge automàtic i, al seu torn, en el funcionament dels sistemes de resolució de la coreferència. A més de l'aspecte quantitatiu, que essencialment es traduïa en un procés costós en temps i diners, el repte de l'anotació residia a guanyar una comprensió a fons del fenomen de la coreferència:

> Fins ara, els models estadístics de resolució de l'anàfora no han fet més que començar a tractar el fenomen de l'anàfora i les contribucions a l'explicació lingüística del fenomen han estat poques. (POESIO *et al.*, en preparació:85)

> Les nocions de coreferència i anàfora són difícils de definir amb precisió i de fer-les operacionals de manera consistent. A més, les connexions entre elles són extremadament complexes. (STOYANOV *et al.*, 2009:657)

En la meva aproximació al problema, vaig dur a terme un estudi de casos empírics i vaig considerar propietats específiques de cada llengua per tal de definir un esquema d'anotació precís (capítol 2). Volia alliberar l'anotació dels requeriments computacionals com els imposats per l'esquema del MUC (HIRSCHMAN i CHINCHOR, 1997), que supedita la definició de la tasca de coreferència a les tasques d'extracció d'informació de la competició MUC. En un estadi posterior, i a un nivell més teòric, em van interessar els reptes de comparar la coreferència i la paràfrasi, un altre fenomen amb el qual guarda molta semblança (capítol 7), i de definir correctament "identitat de referència", una noció que es dóna per entesa però que és

conceptualment molt complexa (capítol 8): Es refereixen a la mateixa entitat *Postville* i *el vell Postville*? *El vas trencat* és el mateix que *la peça sencera*? Aquestes han estat qüestions llargament debatudes i reflectides en les anomenades paradoxes d'identitat com el riu d'Heràclit i el vaixell de Theseus.

Amb un corpus com a estàndard de referència a la meva disposició –anotat no només amb relacions de coreferència sinó també amb informació morfològica, sintàctica i semàntica– em vaig trobar davant el repte de la resolució. D'una banda, pretenia servir-me de l'experiència d'anotació per definir trets motivats lingüísticament (capítol 3). M'interessava explorar l'espai de trets amb l'ajuda de tècniques d'aprenentatge automàtic, conegudes per la seva eficàcia per tractar un gran nombre de trets. D'altra banda, les limitacions dels models basats en parells de mencions podien superar-se dissenyant un sistema basat en entitats que prengués en compte *clusters* sencers d'entitats (capítol 4). D'aquesta manera es podia utilitzar una quantitat major d'informació lingüística per decidir si afegir o no una menció dins d'una entitat ja creada.

Tan aviat com es van obtenir els primers resultats del sistema prototip presentat en aquesta tesi, vaig haver d'assumir el repte de l'avaluació. Tot i que existeix més d'un mètrica per mesurar el funcionament dels sistemes de resolució de la coreferència (VILAIN *et al.*, 1995; BAGGA i BALDWIN, 1998; LUO, 2005), encara no hi ha acord per un estàndard. Calia establir criteris per valorar la qualitat del resultat d'un sistema de coreferència: Què hauria de ser més premiat a (1), lligar com a coreferents les mencions 1, 2 i 3 juntament amb *la vasta glacera*, o bé lligar només les mencions 1 i 3? O, què caldria penalitzar més, lligar la menció 1 i *l'emergència d'un núvol vermell que brillava per sobre de la vasta glacera que el cobreix*, o bé lligar *el capvespre del dia 20 de març de 2010* amb *la vasta glacera*?

Malgrat els molts resultats que han estat publicats, diferents idees que la metodologia d'avaluació dóna per suposades dificulten la comparació entre els resultats dels millors sistemes. Vaig aplicar diferents criteris per analitzar els pros i els contres de les mètriques de coreferència usades actualment (capítol 5). La tasca d'avaluació del SemEval (capítol 6) va ser una eina més per afrontar els reptes presentats per l'avaluació i la comparació de sistemes (capítol 4).

1.2 Esquema de la tesi

La present tesi doctoral reuneix un total de set publicacions precedides d'una introducció general i seguides d'un capítol final de conclusions. El capítol inicial i final donen al conjunt de la tesi la coherència necessària per constituir un tot. Les set publicacions són les següents:

Part I: Anotació de corpus amb coreferència

1. Recasens, Marta i M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Part II: Resolució i avaluació de la coreferència

2. Recasens, Marta i Eduard Hovy. 2009. A deeper look into features for coreference resolution. Dins S. Lalitha Devi, A. Branco, i R. Mitkov (eds.), *Anaphora Processing and Applications (DAARC 2009)*, LNAI 5847:29-42. Springer-Verlag, Berlín.

3. Recasens, Marta i Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. Dins *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, ps. 1423–1432, Uppsala, Suècia.

4. Recasens, Marta i Eduard Hovy. En premsa. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*.

5. Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, i Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. Dins *Proceedings of the ACL 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 1–8, Uppsala, Suècia.

Part III: Teoria de la coreferència

6. Recasens, Marta i Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4).

7. Recasens, Marta, Eduard Hovy, i M. Antònia Martí. En revisió. Identity, non-identity, and near-identity: Addressing the complexity of coreference. Enviat a *Lingua*.

Els primers sis articles han estat publicats, o ho seran aviat, en revistes o actes de congrés amb comitè avaluador d'admissió extern. L'últim article es troba actualment en procés de revisió. Totes les publicacions estan escrites conjuntament amb un o els dos dels meus directors exceptuant-ne dues: l'article 5 feia necessària la col·laboració entre diferents grups de recerca per tal d'organitzar la tasca compartida; l'article 6 va ser el resultat d'un treball conjunt amb una altra doctoranda de la Universitat de Barcelona que investiga la paràfrasi. En tots els casos figuro com la primera autora. S'ha inclòs una còpia completa de les publicacions presentades uniformant-ne el format per aconseguir una tipografia consistent i s'han integrat les referències bibliogràfiques i els apèndixs en una única bibliografia i apèndix al final.

La tesi s'organitza en tres parts. La primera part, formada pel següent capítol, se centra en l'anotació de corpus amb informació sobre la coreferència utilitzant el cas del corpus AnCora del català i castellà. La segona part de la tesi, formada pels capítols 3 al 6, exposa la meva experiència a desenvolupar i avaluar sistemes de resolució de la coreferència. En concret em centro en la definició del conjunt de

trets, la interdependència entre el sistema i els paràmetres del corpus, el comportament de les mètriques d'avaluació i la tasca compartida del SemEval que va establir una plataforma d'avaluació per comparar diferents sistemes. La tercera part de la tesi, que correspon als capítols 7 i 8, ofereix una descripció teòrica més completa de la coreferència: el primer capítol delimita l'abast del concepte contraposant-lo al de paràfrasi i el segon capítol presenta un model de coreferència entesa com a *contínuum* que introdueix la noció de *quasi-identitat* i que obre noves vies de recerca.

En el que resta de capítol, es defineixen primer una sèrie de paraules clau pròpies del camp de la coreferència i es resumeixen els treballs anteriors per posar aquest en perspectiva. Després s'exposa la interconnexió entre les set publicacions detallant la manera en què els resultats de les diferents etapes estan lligats mútualment i van portar-me a emprendre noves línies de recerca. Finalment, es recapitulen les aportacions principals.

1.3 Paraules clau

Per començar, i per tal de clarificar nocions comunament utilitzades en el camp de la resolució de la coreferència, dono una breu explicació i faig algunes puntualitzacions sobre els termes que són rellevants per a aquest estudi.

Menció i **entitat** Els programes MUC i ACE[3] (HIRSCHMAN i CHINCHOR, 1997; DODDINGTON *et al.*, 2004) han popularitzat aquests dos termes en el camp de la resolució de la coreferència. Tal com la defineix l'ACE, una *entitat* és "un objecte o conjunt d'objectes del món". A més, el programa ACE restringeix les entitats a uns determinats tipus específics (persona, organització, lloc, etc.). Una *menció*, d'altra banda, és una referència textual a una entitat. En síntesi, una entitat està constituïda pel conjunt de mencions que s'hi refereixen. Cal fer dues observacions. En primer lloc, les mencions són SNs referencials, per la qual cosa exclouen pronoms expletius (*El nen no li menja*), SNs atributius o predicatius (*És un membre de la companyia*) i SNs idiomàtics (*Plou a bots i barrals*). Alguns plantejaments, tanmateix, adopten una interpretació més àmplia i utilitzen "menció" com a sinònim de "SN". En segon lloc, l'afirmació que les entitats es troben *en el món* ha de ser reconsiderada, tal com es discuteix a continuació.

Model de discurs i **entitat/referent discursiu** Les teories de representació del discurs, que s'ocupen dels processos que intervenen en la comprensió i producció del discurs, consideren que la referència lingüística no és una correspondència entre expressions lingüístiques i el món real sinó entre expressions lingüístiques i els

[3]Els congressos Message Understanding Conferences (MUC) i l'avaluació Automatic Content Extraction (ACE) van ser iniciats i finançats per l'agència DARPA del Departament de Defensa dels Estats Units, i l'Institut Nacional d'Estàndards i Tecnologia dels Estats Units, respectivament, amb el propòsit de fomentar el desenvolupament de nous i millors mètodes d'extracció d'informació.

components del model de discurs que es construeix a mesura que avança el text. En el següent fragment, PRINCE (1981:235) en resumeix les nocions bàsiques:

> Diguem que un text és un conjunt d'instruccions d'un parlant a un interlocutor sobre com construir un *model de discurs* particular. El model contindrà *entitats discursives*, atributs i vincles entre entitats. Una entitat discursiva és un objecte del model de discurs, semblant a un *referent discursiu* de KARTTUNEN (1976); pot representar un individu (existent o no en el món real), una classe d'individus, un exemplar, una substància, un concepte, etc. Segons WEBBER (1979), les entitats es poden pensar com ganxos dels quals es pengen atributs. Totes les entitats discursives d'un model de discurs es representen en el text per SNs, encara que no tots els SNs d'un text representen entitats discursives.

A partir d'ara, sempre que aparegui el terme "entitat" s'ha d'entendre en el sentit d'"entitat discursiva".

Anàfora i **antecedent** S'anomena *anàfora* a una expressió lingüística que depèn d'una expressió anterior (el seu *antecedent*) per a la seva interpretació. El pronom *el* té *Lluís Companys* d'antecedent a *Lluís Companys és homenatjat al poble on el van detenir*. L'anàfora és una relació textual que requereix que el lector vagi endarrera en el text per interpretar un element textual que és buit (o quasi buit), mentre que la coreferència es dóna al nivell referencial. Els termes "anàfora" i "antecedent" són propis del domini de la resolució de l'anàfora, però també s'utilitzen en la resolució de la coreferència de dues maneres: una correcta i una altra incorrecta. S'utilitzen correctament per referir-se a parells d'antecedent-anàfora que són part d'una cadena de coreferència, com *Lluís Companys* i *el*, o *Eyjafjallajökull* i *el volcà* a (1) citat més amunt; mentre que s'utilitzen incorrectament per referir-se indistintament a qualsevol parell de mencions d'una cadena de coreferència, com *el volcà Eyjafjallajökull* i *Eyjafjallajökull*, ja que la segona menció no necessita la primera per ser interpretada.

Primera menció i **menció subsegüent** Els termes homòlegs a "antecedent" i "anàfora" en el camp de la resolució de la coreferència són *primera menció* i *menció subsegüent*, respectivament. Una primera menció introdueix una entitat (nova) en el text; és per tant la seva primera menció, com $Eyjafjallajökull_1$ a (2) més amunt. Una menció subsegüent, en canvi, és qualsevol menció posterior d'una entitat ja introduïda en el text, com $Eyjafjallajökull_2$ i $Eyjafjallajökull_3$ a (2). Noteu que les mencions subsegüents poden ser anafòriques, però no tenen perquè ser-ho. Tanmateix, el terme "no-anafòric" és sovint utilitzat erròniament amb el significat de 'nou en el discurs' i "anafòric" és utilitzat erròniament amb el significat de 'menció subsegüent'.

Entitat unària i **entitat de múltiples mencions** Depenent del tamany de l'entitat, és a dir, del nombre de mencions que conté, és útil distingir entre *entitats unàries* si tenen una única menció i *entitats de múltiples mencions* si es componen de dues o més mencions. Les primeres també s'anomenen mencions *aïllades*, ja que fan una referència aïllada a una entitat. Les relacions de coreferència, doncs, només poden donar-se en entitats de múltiples mencions.

Model basat en parells de mencions i **model basat en entitats** Hi ha dues classes bàsiques de models de resolució de la coreferència. D'una banda, els *models basats en parells de mencions* classifiquen dues mencions o bé com a coreferents o bé com a no coreferents i després combinen totes les decisions de parells per dividir les mencions del document en cadenes de coreferència. D'altra banda, els *models basats en entitats* pretenen millorar la classificació calculant no la probabilitat que una menció corefereixi amb una menció prèvia sinó la probabilitat que una menció es refereixi a una entitat anterior, és a dir, a un conjunt de mencions ja classificades com a coreferents. El segon model, per tant, sol fer ús d'estratègies de *clustering*.

Mesura basada en enllaços i **mesura basada en classes** Paral·lelament a les dues classes de models de resolució, els sistemes de coreferència poden avaluar-se utilitzant dos tipus de mesures: les *mesures basades en enllaços* tenen en compte el número de lligams (és a dir, parells de mencions) identificades pel sistema que són correctes, errònies o que falten, mentre que les *mesures basades en classes* tracten les entitats en termes de *clusters* i tenen en compte no només les entitats de múltiples mencions sinó també les unàries.

Sistema integral i **mòdul de coreferència** El terme *sistema de resolució de la coreferència* és ambigu entre un *sistema integral*, un sistema capaç de determinar la coreferència sobre textos plans, és a dir, d'identificar les mencions i els seus límits automàticament abans de predir relacions de coreferència, i un *mòdul de coreferència*, que es limita a identificar relacions de coreferència pressuposant que les dades contenen informació lingüística de l'estàndard de referència a diferents nivells (límits de les mencions, PoS, arbres sintàctics, etc.).

Mencions reals i **mencions del sistema** A les mencions que conté el corpus produït per experts humans que es pren com a estàndard de referència, se les anomena *mencions reals*, en contraposició al conjunt de mencions que conté la sortida del sistema, que s'anomenen *mencions del sistema*. Per complementar allò dit al paràgraf anterior, les mencions reals i del sistema no acostumen a coincidir quan s'utilitza un sistema integral de coreferència, a diferència de quan es treballa amb un mòdul de coreferència.

1.4 Antecedents

Existeix un bon nombre d'estudis (POESIO *et al.*, en preparació; NG, 2010; NG, 2003) que ofereixen una visió àmplia del tractament computacional de la resolució de la coreferència i l'anàfora. L'objectiu d'aquest apartat és donar al lector no una descripció més, sinó una visió general i compacta de la recerca anterior i recent dels aspectes principals relacionats amb aquesta tesi, per tal de complementar i cohesionar els apartats d'"Antecedents" que ja conté cada publicació.

Des d'una perspectiva temàtica, es pot dividir la matèria objecte d'estudi en les quatre àrees que corresponen aproximadament als quatre passos seguits per desenvolupar un sistema de resolució de la coreferència: (i) creació de corpus, (ii) trets d'aprenentatge automàtic, (iii) classificació i agrupament, i (iv) avaluació. Em referiré a cada punt en aquest mateix ordre, destacant les principals tendències i les fites més importants. Per a més detalls el lector pot consultar les publicacions originals.

1.4.1 Creació de corpus

La investigació en resolució de la coreferència en l'àmbit del PLN requereix corpus anotats amb informació coreferencial per dues raons principals: (i) per entrenar sistemes d'aprenentatge automàtic i (ii) per avaluar els sistemes automàtics. Els corpus anotats amb informació coreferencial són també valuosos per a la lingüística basada en l'ús per estudiar el llenguatge a partir de dades reals. Anotar coreferència, però, no és una tasca trivial i s'han proposat diversos esquemes de codificació. Com que no hi ha cap acord sobre un estàndard, cada corpus sol definir el seu propi esquema. Globalment és possible distingir entre els enfocaments orientats a l'aplicació i els orientats a la lingüística.

Enfocaments orientats a l'aplicació Els corpus MUC i ACE (HIRSCHMAN i CHINCHOR, 1997; DODDINGTON *et al.*, 2004) van ser dissenyats específicament per a tasques compartides en extracció d'informació. En conseqüència, les decisions d'anotació –com el conjunt d'elements coreferents o l'abast de la relació d'identitat– estan subordinades a les necessitats de les tasques, encara que sigui en detriment de la precisió lingüística. Així, no anoten el SN amb tots els seus complements, sinó només fins el nucli, i ignoren les mencions verbals o a nivell de clàusula. ACE també restringeix el conjunt de SNs a set tipus semàntics (els rellevants per al domini de l'extracció d'informació): persona, organització, entitat geopolítica, lloc, instal·lació, vehicle i arma. En la mateixa línia, el MUC i l'ACE adapten la definició d'"identitat de referència" per cobrir les necessitats de l'extracció d'informació i tracten els predicats nominals i les aposicions també com a coreferencials. Per aquesta raó, VAN DEEMTER i KIBBLE (2000) critiquen durament l'esquema del MUC ja que combina "elements genuïns de coreferència amb elements propis de l'anàfora i la predicació d'una forma poc clara i, a vegades, contradictòria".

Enfocaments orientats a la lingüística Com a resposta a la definició del MUC, el meta-esquema MATE, els seus derivats (GNOME, ARRAU [POESIO, 2004*a*; POESIO i ARTSTEIN, 2008]) i altres esquemes com OntoNotes (PRADHAN *et al.*, 2007*b*) tenen l'objectiu de crear corpus no per a una tasca específica sinó per a la investigació en coreferència en general. Aquests esquemes inclouen un ventall més ampli de tipus sintàctics (és a dir, no només els SNs poden ser mencions) i les mencions nominals estan constituïdes per SNs amb tots els seus complements. L'esquema MATE va fins i tot més enllà i contempla fenòmens lingüístics típics de les llengües romàniques, com són els subjectes el·líptics i els clítics incorporats al verb. Es manté una separació estricta entre la relació d'identitat d'una banda i la relació d'aposició (per a predicats nominals i aposicions) per l'altra. MATE també considera l'anotació de relacions diferents de la d'identitat (pertinença a un grup, subconjunt, possessió, anàfora lligada, etc.) així com les ambigüitats.

Tot i la distinció entre aquestes dues direccions, la diversitat d'esquemes existents i les etiquetes idiosincràtiques que incorporen corpus diferents és un reflex de la manca d'una teoria general i plausible de la coreferència no basada únicament en la definició d'"identitat de referència". L'elecció del corpus quan es desenvolupa o avalua un sistema de coreferència no és una qüestió menor i la manera en què els corpus s'han anotat ha determinat en gran mesura el disseny i l'arquitectura dels sistemes.

Recursos lingüístics La taula 1.1 mostra una relació dels corpus existents de més de 25.000 paraules anotats amb informació sobre la coreferència. Predominen els textos periodístics. Per completar la taula, també incloc els dos corpus que constitueixen una de les aportacions d'aquesta tesi (Ancora-Ca i Ancora-Es).[4] Es pot veure clarament que aquests corpus omplen un buit de recursos d'aquest tipus tant per al català com el castellà. Amb anterioritat a AnCora-Ca, no hi havia dades del català anotades amb coreferència i Ancora-Es supera el corpus del castellà de l'ACE-2007 no només en grandària sinó també en les limitacions imposades per les classes d'entitats ACE. Atès que aquesta tesi se centra en la coreferència, la taula 1.1 no inclou els corpus anotats només amb pronoms anafòrics com, per exemple, el corpus del castellà Cast3LB (NAVARRO, 2007). El projecte en curs conegut com ANAWIKI té l'objectiu de recollir grans quantitats de dades per a l'anglès anotades amb coreferència a través d'un joc d'internet anomenat *Phrase Detectives* (POESIO *et al.*, 2008). Encara que la majoria de recursos disponibles són per a l'anglès, destaca l'interès creixent dels últims anys per proveir altres llengües de corpus anotats amb coreferència.

[4]El nom d'*AnCora* (ANnotated CORporA) és un genèric per referir-se als corpus del català i el castellà amb totes les seves capes d'anotació. Per referir-se a una part específica, es pot afegir un sufix al final del nom: *AnCora-Ca* identifica el corpus del català, *AnCora-Es* identifica el corpus del castellà, *AnCora-CO* identifica les anotacions de coreferència dels dos corpus, etc.

Corpus	Cita	Llengua	Gènere	Tamany
ACE-2	DODDINGTON *et al.* (2004)	anglès	notícies	180.000
ACE-2003, ACE-2004, ACE-2005		àrab, xinès, anglès	notícies, weblogs	100-350.000
ACE-2007		àrab	notícies, weblogs	220.000
		xinès	notícies, weblogs	250.000
		anglès	notícies, diàlegs, weblogs, fòrums	300.000
		castellà	notícies	200.000
AnATAr	HAMMAMI *et al.* (2009)	àrab	notícies, llibre de text, novel·la, manual tècnic	77.000
AnCora-Ca	RECASENS i MARTÍ (2010)	català	notícies	400.000
AnCora-Es		castellà	notícies	400.000
ARRAU	POESIO i ARTSTEIN (2008)	anglès	diàlegs, narracions orals, notícies, GNOME	100.000
C-3	NICOLAE *et al.* (2010)	anglès	notícies, tests d'aptitud	75.000
COREA	HENDRICKX *et al.* (2008)	holandès	notícies, discurs oral, entrades d'enciclopèdia	325.000
DAD	NAVARRETTA (2009*b*)	danès, italià	notícies, textos legals, narracions	25.000
GNOME	POESIO (2004*a*)	anglès	rètols de museu, prospectes, diàlegs	50.000
I-CAB	MAGNINI *et al.* (2006)	italià	notícies	250.000
KNACK-2002	HOSTE i DE PAUW (2006)	holandès	notícies	125.000
LiveMemories	RODRÍGUEZ *et al.* (2010)	italià	notícies, Viquipèdia, diàlegs, blogs	150.000
MUC-6	GRISHMAN i SUNDHEIM (1996)	anglès	notícies	30.000
MUC-7	HIRSCHMAN i CHINCHOR (1997)	anglès	notícies	25.000
NAIST Text	IIDA *et al.* (2007)	japonès	notícies	970.000
NP4E	HASLER *et al.* (2006)	anglès	notícies	50.000
Switchboard	CALHOUN *et al.* (2010)	anglès	converses telefòniques	200.000
OntoNotes 2.0	PRADHAN *et al.* (2007*a*)	anglès	notícies	500.000
		àrab	notícies	100.000
		xinès	notícies	550.000
Potsdam Commentary	STEDE (2004)	alemany	notícies	33.000
PDT 2.0	KUČOVÁ i HAJIČOVÁ (2004)	txec	notícies	800.000
TüBa-D/Z	HINRICHS *et al.* (2005)	alemany	notícies	800.000
Venex	POESIO *et al.* (2004*b*)	italià	notícies, diàlegs	40.000

Taula 1.1: Relació dels corpus més grans anotats amb coreferència

1.4.2 Trets d'aprenentatge automàtic

Abans que, a mitjans dels anys noranta, grans quantitats de dades anotades amb relacions de coreferència fossin disponibles, els predecessors immediats dels actuals sistemes de coreferència basats en aprenentatge automàtic van ser els sistemes de resolució de l'anàfora pronominal que es basaven en un conjunt de regles fetes a mà (HOBBS, 1978; RICH i LUPERFOY, 1988; CARBONELL i BROWN, 1988; ALSHAWI, 1990; KAMEYAMA, 1998; TETREAULT, 2001; PALOMAR *et al.*, 2001), especialment en la forma de restriccions i preferències.

Restriccions i preferències Davant d'un pronom que cal resoldre, les restriccions descarten els antecedents que són incompatibles, mentre que les preferències ordenen la resta de candidats per ordre de preferència per seleccionar el millor antecedent. Tant les restriccions com les preferències es basen en informació de diferents nivells lingüístics, tal com mostra la primera fila de la taula 1.2, tot i que posen major èmfasi en la sintaxi (HOBBS, 1978; CARBONELL i BROWN, 1988) i la Teoria del *centering* (KAMEYAMA, 1998; TETREAULT, 2001). Hi va haver, però, una tendència creixent a substituir els sistemes rics en coneixement per sistemes pobres en coneixement que no necessiten informació semàntica o del món (LAPPIN i LEASS, 1994) o, encara més, que no necessiten una anàlisi sintàctica completa (KENNEDY i BOGURAEV, 1996; BALDWIN, 1997; MITKOV, 1998).

Heurístiques d'aquest estil van servir per identificar les regles més importants que regeixen les relacions entre pronom i antecedent. Tanmateix, el major nivell de complexitat de la resolució de la coreferència explica en part la transició de l'última dècada dels mètodes basats en heurístiques als mètodes d'aprenentatge automàtic. L'aplicació de mètodes d'aprenentatge automàtic a gran escala permet ordenar grans conjunts de trets i assignar-los un pes de manera més eficient que els mètodes heurístics basats en regles. Tant AONE i BENNETT (1995) com MCCARTHY i LEHNERT (1995) mostren classificadors basats en aprenentatge que superen els seus homòlegs basats en regles.

Vectors de trets En la configuració clàssica de l'aprenentatge supervisat (vegeu l'apartat 1.4.3), cada exemple d'aprenentatge es crea aparellant dues mencions m_i i m_j i etiquetant-lo o bé com a veritable/coreferent (*exemple positiu*) o bé com a fals/no coreferent (*exemple negatiu*): $\langle m_i, m_j, boolean \rangle$ és veritable si i només si m_i i m_j són coreferents. Els parells $\langle m_i, m_j \rangle$ estan representades per un vector de trets format per trets unaris (és a dir, informació sobre una de les mencions, per exemple, el seu nombre) i trets binaris (és a dir, informació sobre la relació entre les dues mencions, per exemple, concordança de nombre). La taula 1.2 mostra una relació dels trets d'aprenentatge més utilitzats (sobretot per a l'anglès), molts dels quals estan inspirats en restriccions i preferències. Fins ara, la majoria de sistemes de resolució de la coreferència (BENGTSON i ROTH, 2008; DENIS i BALDRIDGE, 2009; STOYANOV *et al.*, 2010) han estat dissenyats seguint el conjunt limitat –però eficaç– de trets de SOON *et al.* (2001), millorat amb l'ampliació de NG i CARDIE

Restriccions i preferències per a la resolució de pronoms (HOBBS, 1978; RICH i LUPERFOY, 1988; CARBONELL i BROWN, 1988; ALSHAWI, 1990; LAPPIN i LEASS, 1994; KENNEDY i BOGURAEV, 1996; BALDWIN, 1997; MITKOV, 1998; KAMEYAMA, 1998; TETREAULT, 2001)	**1.** Concordança de gènere, **2.** Concordança de nombre, **3.** Restriccions de lligament, **4.** m_i és un subjecte, **5.** Animacitat, **6.** Restriccions de selecció, **7.** La menció és a l'interior d'una estructura quantificada o negativa, **8.** Paral·lelisme de cas-rol, **9.** Paral·lelisme sintàctic, **10.** m_i és a l'interior d'una estructura topicalitzada, **11.** Distància en nombre d'oracions, **12.** Novetat, **13.** Funció gramatical, **14.** Concordança de persona, **15.** Freqüència de la menció, **16.** Les postcondicions de l'acció que conté m_i violen les precondicions de l'acció que conté m_j, **17.** La menció és dins d'una altra, **18.** La menció és dins d'una construcció existencial, **19.** Restriccions de *centering*, **20.** m_i és definida, **21.** m_i és el primer SN de l'oració, **22.** m_i és el subjecte de verbs com *discutir, presentar, il·lustrar, descriure*, etc., **23.** m_i és dins del títol de la secció, **24.** La menció no és part d'un sintagma preposicional, **25.** La menció és un terme tècnic [...]
Conjunt bàsic de trets de coreferència (SOON *et al.*, 2001)	**1.** m_i és un pronom, **2.** m_j és un pronom, **3.** m_j és definida, **4.** m_j és un demostratiu, **5.** m_i i m_j són noms propis, **6.** Mateixa cadena (sense els determinants), **7.** Concordança de nombre, **8.** Concordança de gènere, **9.** Concordança de classe semàntica, **10.** m_j és una aposició de m_i, **11.** Una menció és un àlies de l'altra, **12.** Distància en nombre d'oracions
Conjunt ampliat de trets de coreferència (NG i CARDIE, 2002*b*)	**1.** m_i i m_j són pronoms/noms propis/no-pronoms i són la mateixa cadena, **2.** Hi ha intersecció entre els mots de m_i i m_j, **3.** Els complements prenominals d'una menció són un subconjunt dels de l'altra, **4.** m_i i m_j són noms propis/no-pronoms i una és una subcadena de l'altra, **5.** m_i i m_j són definits, **6.** m_i i m_j estan a l'interior d'un SN, **7.** m_i i m_j són part d'una cadena de discurs directe, **8.** m_i és un subjecte, **9.** m_j és un subjecte, **10.** m_i i m_j són subjectes, **11.** m_i i m_j tenen la mateixa animacitat, **12.** m_i i m_j tenen la mateixa projecció màxima de SN, **13.** m_j és un predicat nominal de m_i, **14.** m_i és indefinit i no una aposició, **15.** m_i i m_j no són noms propis però contenen noms propis diferents, **16.** m_i i m_j estan relacionats a WordNet com avantpassat-descendent, **17.** Distància a WordNet, **18.** Distància en nombre de paràgrafs [...]
Trets addicionals (STRUBE *et al.*, 2002; LUO *et al.*, 2004; NICOLAE i NICOLAE, 2006; PONZETTO i STRUBE, 2006; URYUPINA, 2006; NG, 2007; YANG i SU, 2007; BENGTSON i ROTH, 2008)	**1.** Distància mínima d'edició entre les cadenes de m_i i m_j, **2.** Mateix nucli, **3.** Distància en nombre de paraules, **4.** Distància en nombre de mencions, **5.** Una menció és un acrònim de l'altra, **6.** El parell de les cadenes de mencions, **7.** Nombre de paraules en majúscula diferents en les dues mencions, **8.** Rol semàntic, **9.** Valor de semblança a WordNet per a tots els parells de *synsets* de m_i i m_j, **10.** El primer paràgraf de la pàgina de la Viquipèdia titulada m_i conté m_j (o viceversa), **11.** La pàgina de la Viquipèdia titulada m_i conté un hiperenllaç a la pàgina de la Viquipèdia titulada m_j (o viceversa), **12.** Prominència, **13.** Una menció és un sinònim/antònim/hiperònim de l'altra a WordNet, **14.** m_i i m_j apareixen a menys de dues paraules d'un verb de dicció, **15.** Mateixos complements, **16.** Relació dels complements alineats, **17.** Semblança semàntica, **18.** Camí en l'arbre sintàctic de m_j a m_i [...]
Trets a nivell de *cluster* (LUO *et al.*, 2004; DAUMÉ III i MARCU, 2005; NG, 2005; CULOTTA *et al.*, 2007; POON i DOMINGOS, 2008; YANG *et al.*, 2008; RAHMAN i NG, 2009)	**1.** El tret X és veritat per a qualsevol parell, **2.** Tots els parells comparteixen un tret X, **3.** La majoria de parells comparteixen un tret X, **4.** El tret X és fals per a qualsevol parell, **5.** Tots els parells de mencions són predits com a coreferents, **6.** La majoria de parells de mencions són predits com a coreferents, **7.** Densitat de decadència, **8.** Raó d'entitats i mencions, **9.** Tamany de la cadena hipotètica d'entitats, **10.** Nombre de SNs de cada tipus de menció, **11.** Probabilitat que un parell tingui valors de gènere incompatibles [...]

Taula 1.2: Relació dels trets de coreferència més influents (m_i i m_j representen dues mencions diferents on $i < j$)

(2002*b*). Un dels trets que ha resultat obtenir millors resultats és el d'aposició (SOON *et al.*, 2001; POON i DOMINGOS, 2008).

Trets addicionals Encara que els trets superficials i morfosintàctics mostrats a la taula 1.2 aconsegueixen resoldre una gran part de les relacions de coreferència, tenen un límit, sobretot en el cas dels SNs definits i noms propis (VIEIRA i POESIO, 2000; HAGHIGHI i KLEIN, 2009), que només pot ser superat utilitzant coneixement profund de tipus semàntic i del món. Els models més recents han intentat oferir una aproximació útil a aquest coneixement mitjançant l'elaboració de patrons semàntics a partir de recursos com WordNet i internet (PONZETTO i STRUBE, 2006; URYUPINA, 2006; NG, 2007), però qualsevol millora, encara que sigui significativa, és petita. En aquesta línia, KEHLER *et al.* (2004) assenyalen que les estadístiques referents a predicats i arguments extretes de dades reals no milloren els resultats (per a la resolució de pronoms), ja que els casos en què les estadístiques perjudiquen són potencialment més nocius que aquells en què ajuden.

Trets a nivell de *cluster* Una manera prometedora d'incorporar més coneixement sense necessitat de construir recursos costosos és el disseny de models més globals amb trets a nivell de *cluster* que permetin tenir en compte no només dues sinó totes les mencions d'una entitat (parcial) (LUO *et al.*, 2004; CULOTTA *et al.*, 2007; POON i DOMINGOS, 2008). El disseny d'aquests trets, però, és complex i la majoria solen derivar directament dels clàssics trets a nivell de parell. A més d'introduir noves fonts de coneixement per enfortir el conjunt de trets, els resultats també poden millorar significativament mitjançant la selecció de trets (NG i CARDIE, 2002*b*; HOSTE, 2005) i la selecció d'exemples d'entrenament (HARABAGIU *et al.*, 2001; NG i CARDIE, 2002*b*; URYUPINA, 2004; HOSTE, 2005), encara que aquests aspectes han rebut menys atenció. En gran mesura el centre d'interès s'ha desplaçat de la incorporació de nous trets a l'aplicació de nous models de resolució, com es discuteix a 1.4.3.

Llengües diferents de l'anglès A mesura que llengües diferents de l'anglès han disposat de corpus anotats amb coreferència, s'ha pogut comprovar la validesa dels trets de coreferència per a aquestes llengües, però són treballs força recents. Vegeu, per exemple, HOSTE (2005) per a l'holandès, VERSLEY (2007) o KLENNER i AILLOUD (2009) per a l'alemany, POESIO *et al.* (2010) per a l'italià i NILSSON (2010) per al suec. Amb anterioritat a la recerca recollida en aquesta tesi, el cas del català i el castellà romania gairebé inexplorat, llevat d'uns pocs sistemes de resolució pronominal basats en regles per al castellà (PALOMAR *et al.*, 2001; FERRÁNDEZ *et al.*, 1999).

1.4.3 Classificació i agrupament

Per al MUC-6 i MUC-7 es van construir alguns sistemes de coreferència basats en regles (APPELT *et al.*, 1995; GAIZAUSKAS *et al.*, 1995; GARIGLIANO *et al.*, 1997),

però el fet que aquestes competicions duguessin a terme avaluacions a gran escala i que posessin a l'abast una considerable quantitat de dades anotades per a aquest fi va contribuïr significativament en l'ús creixent de tècniques d'aprenentatge automàtic per a la tasca de coreferència. Des de llavors, la investigació en sistemes de resolució de la coreferència ha anat en augment i és precisament en les seves estratègies de resolució que ara em centro. També, però, s'han seguit desenvolupant sistemes de resolució pronominal (YANG *et al.*, 2004; NAVARRETTA, 2004; KEHLER *et al.*, 2004; HINRICHS *et al.*, 2007), especialment per a models computacionals del diàleg (STRUBE i MÜLLER, 2003; FRAMPTON *et al.*, 2009). Els resultats obtinguts per cada sistema es mostren a la taula 1.3 i es discutiran a 1.4.4.

Dues etapes Mitjançant la informació lingüística codificada en els trets resumits a 1.4.2, s'han dissenyat diferents models amb l'objectiu d'agrupar les mencions d'un document en un conjunt d'entitats. La manera en què SOON *et al.* (2001) ha formulat la tasca, inspirada pels primers sistemes d'AONE i BENNETT (1995) i MCCARTHY i LEHNERT (1995), s'ha pres com a referent i constitueix el punt de partida per construir qualsevol sistema de coreferència. D'acord amb aquesta concepció, la tasca de coreferència consisteix en un procediment de dues etapes:

1. Una etapa de *classificació* que decideix si dues mencions corefereixen o no. Es tracta d'un problema de classificació binari en què la probabilitat que la menció m_i i la menció m_j siguin coreferents pot ser calculada estimant la probabilitat que:

$$P_c(m_i, m_j) = P(COREFERENT|m_i, m_j)$$

2. Una etapa d'*agrupament* que converteix el conjunt de classificacions de parells en *clusters* de mencions, creant un *cluster* per a cada entitat. Aquesta etapa necessita coordinar les possibles decisions contradictòries de classificació coreferencial de la primera etapa.

Els sistemes de coreferència poden variar segons ambdues dimensions independentment. En la fase de classificació, la probabilitat de coreferència de dues mencions es pot predir entrenant diferents algorismes d'aprenentatge automàtic: arbres de decisió (SOON *et al.*, 2001; NG i CARDIE, 2002*b*; NG, 2005; YANG i SU, 2007), classificadors de màxima entropia (LUO *et al.*, 2004; NG, 2005; NICOLAE i NICOLAE, 2006; PONZETTO i STRUBE, 2006), inductors de regles RIPPER (NG i CARDIE, 2002*b*; HOSTE, 2005; NG, 2005), màquines de vectors de suport (URYUPINA, 2007; RAHMAN i NG, 2009), aprenentatge basat en memòria (KLENNER i AILLOUD, 2009; HOSTE, 2005) o perceptrons (BENGTSON i ROTH, 2008; STOYANOV *et al.*, 2009). En la fase d'agrupació, a grans trets podem distingir entre els models locals i els models globals o, en altres paraules, entre els models basats en parells de mencions i els models basats en entitats.

Models basats en parells de mencions Els models basats en parells de mencions poden seguir diverses estratègies com *primer-enllaç* (SOON *et al.*, 2001; STRUBE *et al.*, 2002) i *millor-enllaç*, la més utilitzada (NG i CARDIE, 2002*b*; YANG i SU, 2007; BENGTSON i ROTH, 2008). La primera compara cada menció amb cada menció precedent, de dreta a esquerra, i el procés s'acaba tan aviat com s'arriba al principi del text o el classificador retorna una probabilitat de coreferència per sobre de 0,5 per a un parell de mencions, cas en què les dues mencions s'agrupen en una mateixa entitat. En canvi, l'estratègia de millor-enllaç calcula la probabilitat de totes les mencions que precedeixen la menció que s'analitza i selecciona la menció amb la probabilitat de coreferència més elevada (per sobre de 0,5), és a dir, pren la decisió més segura.

Tant primer-enllaç com millor-enllaç són models basats en parells de mencions que presenten el greu inconvenient de ser optimitzats només localment. Com que la coreferència és una relació transitiva,[5] aquests models simplement realitzen el tancament transitiu de les decisions de parells, però no garanteixen la coherència global de l'entitat. Per exemple, *el Sr. Clinton* pot ser correctament coreferit amb *Clinton*, però més endavant certs trets de parells poden fer que el model cregui erròniament que *Clinton* és coreferent amb una menció pròxima d'*ella* i, com que l'etapa d'agrupament és independent de la de classificació de parells, la incompatibilitat entre el gènere d'*el Sr. Clinton* i el d'*ella* passarà desapercebut quan es construeixi l'entitat final. Aquesta és la diferència que separa els models locals dels models globals o basats en entitats.

Models basats en entitats A diferència dels models basats en parells de mencions, els basats en entitats aprofiten la informació proporcionada per altres mencions d'una entitat anterior (i parcial). Això pot ser especialment útil quan és difícil jutjar si dues mencions són o no coreferents a partir del parell sol i pot servir per recuperar un lligam no detectat o evitar-ne un d'erroni. Naturalment, qualsevol sistema que funcioni només amb enllaços (parells) no pot aplicar les restriccions de transitivitat. Amb aquesta finalitat, els últims sistemes de coreferència treballen amb *clusters* que permeten avaluar el grau de compatibilitat entre una menció i una entitat *com un tot*.

Un dels primers sistemes que va avançar en aquesta direcció va ser el de LUO *et al.* (2004), on es consideren totes les possibilitats d'agrupació (és a dir, les particions d'entitats) buscant en un arbre de Bell i es planteja la resolució de la coreferència com el problema de trobar el millor camí des del node arrel fins a les fulles (on cada fulla és una partició possible). Les diferents hipòtesis de partició es construeixen utilitzant un classificador estàndard basat en parells de mencions o bé un classificador basat en entitats, que determina la probabilitat que una menció faci referència a una determinada entitat. Sorprenentment, però, el primer obté resultats més favorables que el segon. NICOLAE i NICOLAE (2006) sostenen que "tot

[5]Per la propietat transitiva, si una menció *a* és coreferent amb *b* i *b* és coreferent amb *c*, llavors *a* és coreferent amb *c*.

i que l'arbre de Bell és una representació completa de l'espai de cerca, la cerca és optimitzada en tamany i temps amb la possibilitat de perdre solucions òptimes." A més, LUO *et al.* (2004) al·ludeixen al menor nombre de trets (vint vegades menys) utilitzats pel model basat en entitats com una raó possible per explicar els resultats més baixos.

Des que LUO *et al.* (2004) va proposar fer una cerca global en un arbre de Bell, s'han suggerit altres formes d'optimitzar globalment la decisió d'agrupament: un model probabilístic de primer ordre que permet definir trets basats en lògica de primer ordre per a un conjunt de mencions (CULOTTA *et al.*, 2007); programació lineal entera per imposar la restricció de transitivitat (FINKEL i MANNING, 2008; KLENNER i AILLOUD, 2009); un algorisme de tall sobre un graf en què els vèrtexs representen mencions i les arestes reben un pes segons la probabilitat de coreferència dels parells (NICOLAE i NICOLAE, 2006; NG, 2009); un model de grafs entrenat condicionalment (MCCALLUM i WELLNER, 2005; WICK i MCCALLUM, 2009); un model d'aprenentatge en línia que aprèn l'estratègia de cerca òptima per si mateix (DAUMÉ III i MARCU, 2005); o programació lògica inductiva per aprendre del coneixement relacional d'una menció, una entitat i les mencions de l'entitat amb un conjunt de regles de primer ordre (YANG *et al.*, 2008). Tot i que aquests models garanteixen la coherència global, no tots inclouen trets a nivell de *cluster* (MCCALLUM i WELLNER, 2005). El disseny d'aquest tipus de trets ha estat poc explorat.

Models no supervisats Entre els sistemes optimitzats globalment hi trobem els pocs sistemes no supervisats que existeixen per a la resolució de la coreferència: HAGHIGHI i KLEIN (2007) fan servir un model generatiu no paramètric bayesià basat en un procés jeràrquic de Dirichlet, i POON i DOMINGOS (2008) introdueixen relacions entre mencions com l'aposició i els predicats nominals adaptant un algorisme d'aprenentatge no supervisat a la lògica de Markov. Tant HAGHIGHI i KLEIN (2007) com POON i DOMINGOS (2008) imposen com a condició prèvia el nombre de *clusters*, que no és el cas del sistema generatiu de NG (2008). NG (2008) modifica l'algorisme esperança-maximització (EM) perquè no calgui predeterminar el nombre de *clusters* i redefineix el pas E per calcular les *n* particions de coreferència més probables utilitzant un arbre de Bell (LUO *et al.*, 2004). El sistema no supervisat i generatiu de HAGHIGHI i KLEIN (2010) aborda el problema de la compatibilitat semàntica entre nuclis mitjançant un gran inventari dels tipus d'entitats distribucionals. Finalment, un dels primers sistemes, el de CARDIE i WAGSTAFF (1999), es troba a cavall de l'aprenentatge supervisat i el no supervisat. Aplica *clustering* en vectors de trets que representen mencions amb l'objectiu de crear un *cluster* per a cada entitat, però no és completament no supervisat perquè la mètrica de distància utilitzada per a la comparació fa servir pesos fixos que es decideixen heurísticament.

Models de rànquing L'estratègia de rànquing es pot considerar com un pas intermedi entre els models locals i els globals. El rànquing permet examinar simultàniament més d'un candidat de menció i, en determinar quin candidat és més probable, recull directament la competència entre ells. El primer model de rànquing (CONNOLLY *et al.*, 1994), que classifica dos candidats (un positiu i un negatiu) a la vegada, és utilitzat per YANG *et al.* (2003) sota el nom de *model de doble candidat* i per IIDA *et al.* (2003) sota el nom de *model de torneig* per a l'anàfora zero en japonès. Els candidats, però, es comparen en forma de parells. En canvi, el rànquing de DENIS i BALDRIDGE (2008) considera tot el conjunt de candidats a la vegada. NG (2005) fa un ús diferent del rànquing i planteja la tasca de coreferència com el problema d'ordenar particions de candidats generades per diferents sistemes basats en parells de mencions. D'aquesta manera pot beneficiarse de les fortaleses dels diferents mètodes. No és un enfocament realment global, però, ja que les particions de candidats són totes generades per models basats en parells de mencions. Finalment, RAHMAN i NG (2009) proposen un enfocament amb *clusters* i rànquing que combina els punts forts dels models basats en entitats i els models de rànquing.

Estratègies potenciadores Altres mètodes que s'han utilitzat per millorar la resolució de la coreferència són separar els mòduls de resolució de pronoms, noms propis i SNs lèxics (MORTON, 2000; MÜLLER *et al.*, 2002; HOSTE, 2005; NG, 2005; HAGHIGHI i KLEIN, 2007; LUO, 2007; DENIS i BALDRIDGE, 2008) així com determinar explícitament la probabilitat que una menció sigui nova en el discurs, ja sigui mitjançant un classificador en una configuració en cascada (NG i CARDIE, 2002*a*; BEAN i RILOFF, 1999; VIEIRA i POESIO, 2000; URYUPINA, 2003; YANG *et al.*, 2003; KABADJOV, 2007; GUODONG i FANG, 2009) o coordinant conjuntament les probabilitats de coreferència i que una menció sigui nova en el discurs (LUO, 2007; NG, 2009), incloent la inferència conjunta o l'aprenentatge (DENIS i BALDRIDGE, 2007; POON i DOMINGOS, 2008; RAHMAN i NG, 2009).

El procés d'aprenentatge també pot ser impulsat filtrant els pronoms expletius (EVANS, 2000; BOYD *et al.*, 2005; BERGSMA *et al.*, 2008) i els SNs indefinits no referencials (BYRON i GEGG-HARRISON, 2004). Per al danès, NAVARRETTA (2009*a*) desenvolupa un classificador automàtic de pronoms neutres i demostratius en múltiples funcions: no referencial, catàforic, díctic, anafòric amb un SN com a antecedent, anafòric amb una oració o clàusula com a antecedent, vague (és a dir, l'antecedent està implícit en el discurs), etc.

Passaré ara a comentar els resultats obtinguts pels diferents sistemes d'aprenentatge discutits fins aquí i plantejaré el problema de l'avaluació. Observeu que la taula 1.3 del següent apartat també inclou el sistema de HAGHIGHI i KLEIN (2009), que és un cas aïllat en els últims anys de l'enfocament basat en regles. Aconsegueix una puntuació comparable a la dels sistemes d'aprenentatge d'última generació tot i que només fa servir un conjunt limitat de restriccions sintàctiques i semàntiques

(per exemple, mateix nucli, concordança i aposició). Contrasta clarament amb alguns dels complexos algorismes d'aprenentatge que han estat implementats.

1.4.4 Avaluació

A l'igual que altres tasques de PLN, avaluar un sistema de resolució de la coreferència implica no només avaluar el seu rendiment, sinó també sospesar els beneficis generals que aporta a l'estat de l'art. Quantificar el funcionament d'un sistema no és senzill. BYRON (2001) i MITKOV i HALLETT (2007) paren atenció a les inconsistències que es detecten en comunicar els resultats de la tasca de resolució pronominal. Assenyalen que els estudis varien respecte de les classes de pronoms tractades i que la majoria dels algorismes es basen en resultats que han estat posteditats. Ambdós factors tenen un efecte evident en la precisió i la cobertura estàndards. Problemes similars, i pitjors encara, es troben en avaluar sistemes de coreferència.

Mencions reals i del sistema En la resolució de la coreferència, una dificultat important per definir una mètrica adequada prové del desconeixement inicial del nombre total d'entitats. Això s'agreuja pel fet que les mencions resoltes pel sistema (*mencions del sistema*), si es detecten automàticament, poden no coincidir amb les mencions de l'estàndard de referència (*mencions reals*). D'altra banda, les diferents mencions considerades per diferents esquemes d'anotació (per exemple, el MUC només anota mencions d'entitats de múltiples mencions i l'ACE només anota mencions de certs tipus semàntics) repercuteixen directament en la complexitat de resoldre un text específic. En conseqüència, els resultats tendeixen a variar considerablement entre corpus diferents (STOYANOV *et al.*, 2009). Aquesta és la raó per la qual la taula 1.3 desglossa les puntuacions segons les dades de test utilitzades.

Mètriques actuals De la mateixa manera que el programa MUC es pot considerar com el punt de partida dels sistemes de resolució de la coreferència i dels corpus anotats amb coreferència a gran escala, va ser també el primer a definir una mètrica d'avaluació, coneguda com la mètrica MUC (VILAIN *et al.*, 1995). Malgrat l'ampli ús que s'ha fet i encara es fa d'aquesta mètrica, se n'han detectat nombroses deficiències en diverses ocasions (BAGGA i BALDWIN, 1998; NG, 2005; NICOLAE i NICOLAE, 2006; DENIS i BALDRIDGE, 2008; FINKEL i MANNING, 2008) i s'han proposat mètriques alternatives, entre les quals B^3 (BAGGA i BALDWIN, 1998) i CEAF (LUO, 2005) segueixen sent les més utilitzades. Aquestes mesures es discuteixen amb més detall al capítol 5, però en sintetitzo aquí les fórmules:[6]

[6]Cada mètrica es calcula en termes de cobertura (C), una mesura d'exhaustivitat, i precisió (P), una mesura d'exactitud. La mesura F correspon a la mitjana harmònica: $F = 2 \cdot P \cdot C / (P + C)$.

- Mètrica MUC

$$\mathrm{C} = \frac{\text{\# enllaços en comú entre la partició real i la del sistema}}{\text{\# enllaços mínims per a la partició real}}$$

$$\mathrm{P} = \frac{\text{\# enllaços en comú entre la partició real i la del sistema}}{\text{\# enllaços mínims per a la partició del sistema}}$$

- B^3

$$\mathrm{C} = \frac{\sum_{i=1}^{n} \frac{\text{\# mencions en comú entre l'entitat de la menció real i la del sistema}_i}{\text{\# mencions dins l'entitat de la menció real}_i}}{n}$$

$$\mathrm{P} = \frac{\sum_{i=1}^{n} \frac{\text{\# mencions en comú entre l'entitat de la menció real i la del sistema}_i}{\text{\# mencions dins l'entitat de la menció del sistema}_i}}{n}$$

- CEAF-ϕ_3

$$\mathrm{C/P} = \frac{\text{\# mencions en comú entre les entitats reals i les del sistema millor alineades}}{\text{\# mencions dins la partició real/del sistema}}$$

La mètrica B^3 va ser dissenyada per corregir les dues deficiències més grans de la mesura MUC: la preferència pels sistemes que creen entitats grans i el fet d'ignorar les entitats unàries correctament detectades. La mètrica CEAF, al seu torn, va ser proposada per resoldre una debilitat de B^3: el fet que una entitat pot ser utilitzada més d'una vegada quan s'alineen les mencions reals amb les del sistema. Malgrat els inconvenients reconeguts de la mètrica MUC, s'ha seguit utilitzant per dues raons: (i) a efectes de comparació amb els sistemes més antics que només estan documentats amb la mètrica MUC, i (ii) a causa de la manca d'acord sobre una mesura estàndard ja que cap ha demostrat ser clarament superior. Actualment, doncs, l'avaluació de sistemes de coreferència es fa donant els resultats en una o dues o les tres mètriques, com mostra la taula 1.3.

Resultats actuals La mètrica MUC és l'única per a la qual tenim les puntuacions de gairebé tots els sistemes, però això no és molt útil ja que emetre judicis de qualitat que recolzen en una mètrica defectuosa (vegeu els arguments del capítol 5) seria enganyós per raons òbvies. DENIS i BALDRIDGE (2008) són partidaris que els resultats de coreferència no es presentin únicament en termes de MUC. La incompatibilitat entre els rànquings de sistemes que s'obtenen segons MUC i B^3 (compareu, per exemple, l'antepenúltima fila i la prèvia a la taula 1.3) és una prova més que no és possible confiar solament en la mètrica MUC. El problema és que alguns sistemes es mesuren o bé amb B^3 o bé amb CEAF, resultats que no són comparables.

En conjunt, només és raonable concloure que no hi ha un guanyador clar per a l'estat de l'art. La manca d'una mètrica fiable, l'ús de corpus diferents (i de porcions diferents del mateix corpus) i la dependència en els límits de mencions

Sistema	Mencions	MUC			B^3			CEAF		
		P	C	F	P	C	F	P	C	F
Dades del MUC-6										
CARDIE i WAGSTAFF (1999)	Sistema	54,6	52,7	53,6						
MORTON (2000)	Sistema	79,6	44,5	57,1						
HARABAGIU *et al.* (2001)	Reals	92	73,9	81,9						
SOON *et al.* (2001)	Sistema	67,3	58,6	62,6						
NG i CARDIE (2002*b*)	Sistema	78,0	64,2	70,4						
YANG *et al.* (2003)	Reals	80,5	64,0	71,3						
LUO *et al.* (2004)	Reals			85,7						76,8
MCCALLUM i WELLNER (2005)	Reals	80,5	64,0	71,3						
CHOI i CARDIE (2007)	Sistema	69,3	70,5	69,9						
HAGHIGHI i KLEIN (2007)	Reals	80,4	62,4	70,3						
FINKEL i MANNING (2008)	Reals	89,7	55,1	68,3	90,9	49,7	64,3			
POON i DOMINGOS (2008)	Reals	83,0	75,8	79,2						
HAGHIGHI i KLEIN (2009)	Reals	87,2	77,3	81,9	84,7	67,3	75,0	72,0	72,0	72,0
STOYANOV *et al.* (2009)	Sistema			68,5			70,9			
Dades del MUC-7										
SOON *et al.* (2001)	Sistema	65,5	56,1	60,4						
NG i CARDIE (2002*b*)	Sistema	70,8	57,4	63,4						
YANG *et al.* (2003)	Reals	75,4	50,1	60,2						
URYUPINA (2007)	Sistema	67,0	50,5	65,4						
STOYANOV *et al.* (2009)	Sistema			62,8			65,9			
Dades de l'ACE-2										
LUO *et al.* (2004)	Reals									73,1
NICOLAE i NICOLAE (2006)	Reals	91,1	88,2	89,6				82,7	82,7	82,7
	Sistema	52,0	82,4	63,8						41,2
DENIS i BALDRIDGE (2007)	Reals	77,1	63,6	69,7						
YANG i SU (2007)	Sistema	73,9	56,5	64,0						
DENIS i BALDRIDGE (2008)	Reals	75,7	67,9	71,6	79,8	66,8	72,7	67,0	67,0	67,0
FINKEL i MANNING (2008)	Reals	83,3	52,7	64,1	90,2	62,6	73,8			
POON i DOMINGOS (2008)	Reals	68,4	68,5	68,4	71,7	66,9	69,2			
NG (2009)	Sistema	69,2	55,0	61,3				59,6	63,7	61,6
STOYANOV *et al.* (2009)	Sistema			66,0			78,3			
Dades de l'ACE-2003										
PONZETTO i STRUBE (2006)	Sistema	84,2	61,0	70,7						
NG (2008)	Reals	69,9	51,6	59,3				61,1	61,1	61,1
	Sistema	63,3	48,8	54,7				55,6	60,0	57,7
YANG *et al.* (2008)	Sistema	60,5	63,4	61,8						
STOYANOV *et al.* (2009)	Sistema			67,9			79,4			
Dades de l'ACE-2004										
LUO i ZITOUNI (2005)	Reals							82,0	82,0	82,0
CULOTTA *et al.* (2007)	Reals				86,7	73,2	79,3			
HAGHIGHI i KLEIN (2007)	Reals	65,0	61,8	63,3						
BENGTSON i ROTH (2008)[a]	Reals	82,7	69,9	75,8	88,3	74,5	80,8			
POON i DOMINGOS (2008)	Reals	69,1	69,2	69,1						
HAGHIGHI i KLEIN (2009)[a]	Reals	74,8	77,7	79,6	79,6	78,5	79,0	73,3	73,3	73,3
	Sistema	67,5	61,6	64,4	77,4	69,4	73,2			
STOYANOV *et al.* (2009)[a]	Sistema			62,0			76,5			
WICK i MCCALLUM (2009)	Reals	78,1	63,7	70,1	87,9	76,0	81,5			
HAGHIGHI i KLEIN (2010)[a]	Sistema	67,4	66,6	67,0	81,2	73,3	77,0			
Dades de l'ACE-2005										
LUO (2007)	Reals							84,8	84,8	84,8
RAHMAN i NG (2009)	Reals	83,3	69,9	76,0	74,6	56,0	64,0	63,3	63,3	63,3
	Sistema	75,4	64,1	69,3	70,5	54,4	61,4	62,6	56,7	59,5
STOYANOV *et al.* (2009)	Sistema			67,4			73,7			
HAGHIGHI i KLEIN (2010)[b]	Sistema	74,6	62,7	68,1	83,2	68,4	75,1			
HAGHIGHI i KLEIN (2010)[c]	Sistema	77,0	66,9	71,6	55,4	74,8	63,8			

[a] Dades de test de l'ACE-2004 utilitzades a CULOTTA *et al.* (2007).
[b] Dades de test de l'ACE-2005 utilitzades a STOYANOV *et al.* (2009).
[c] Dades de test de l'ACE-2005 utilitzades a RAHMAN i NG (2009).

Taula 1.3: Relació dels resultats més destacats de sistemes de resolució de la coreferència

reals o del sistema (per exemple, a diferència de HARABAGIU *et al.* [2001], SOON *et al.* [2001] i NG i CARDIE [2002*b*] parteixen de conjunts de dades sense cap mena de preprocessament) fan que no tingui sentit comparar els sistemes entre ells. Dur a terme una avaluació qualitativa, d'altra banda, només és possible per als pocs sistemes que s'han fet públics:[7] les sortides d'aquests sistemes acostumen a lligar un nombre elevat de sintagmes no coreferents com a coreferents i viceversa.

1.5 Fil conductor

Els quatre aspectes de la coreferència que aquesta tesi tracta –teoria, anotació, resolució i avaluació– estan íntimament relacionats però són separables a la vegada. Per raons d'espai i claredat, em limito essencialment a un aspecte del problema en cada publicació, però sempre posant-lo en perspectiva sobre com afecta i està connectat amb la resta del problema. Aquest apartat integra les diferents perspectives. Discuteixo el marc general, les troballes dels articles que posen de manifest una sèrie d'indicis de ruptura, i les noves vies que he iniciat per satisfer les necessitats de la tasca de resolució de la coreferència.

1.5.1 Metodologia

Des de bon principi, vaig enfocar el problema de la coreferència des d'una metodologia basada en corpus i per això he donat prioritat a l'ús de dades reals. La meva preocupació principal era estudiar la coreferència tal com es dóna en la llengua real. Així, doncs, molts dels problemes plantejats en aquesta tesi queden fora de l'abast de la lingüística teòrica i la psicolingüística, els estudis de les quals acostumen a estar restringits a exemples que s'han construït atentament o que s'han seleccionat amb molta cura. Com que aquests exemples rarament tenen més de dues frases, no contenen les relacions típiques de discursos llargs escrits. En concret, la meva investigació s'ha centrat en textos periodístics utilitzant un total de sis corpus:

AnCora (RECASENS i MARTÍ, 2010) Un banc d'arbres sintàctics del català i el castellà amb 500.000 paraules cadascun format majoritàriament per textos de diaris i d'agències de notícies (El Periódico, EFE, ACN). A part de la morfosintaxi, està anotat manualment amb arguments i rols temàtics , classes semàntiques dels predicats, entitats amb nom, sentits nominals de WordNet i relacions de coreferència (anotació que és part del treball d'aquesta tesi).

ACE (DODDINGTON *et al.*, 2004) El conjunt de dades de l'anglès per als programes ACE 2003, 2004 i 2005 inclou notícies de diaris, agències i noticiaris de la col·lecció TDT. Estan anotades amb els tipus d'entitats de l'ACE (persona, organització, lloc, entitat geopolítica, instal·lació, etc.), subtipus d'entitat,

[7] Entre els sistemes de coreferència disponibles gratuïtament hi ha OpenNLP (http://opennlp.sourceforge.net), BART (VERSLEY *et al.*, 2008), Reconcile (STOYANOV *et al.*, 2010), Illinois Coreference Package (BENGTSON i ROTH, 2008), CoRTex (DENIS i BALDRIDGE, 2008), CherryPicker (RAHMAN i NG, 2009) i ARKref (http://www.ark.cs.cmu.edu/ARKref).

classe de menció (específica, genèrica, atributiva, etc.) i les mencions d'una mateixa entitat estan agrupades. Els corpus han estat creats i són distribuïts pel Linguistic Data Consortium.

OntoNotes (PRADHAN *et al.*, 2007*a*) El corpus OntoNotes de l'anglès, versió 2.0, conté notícies i noticiaris: 300.000 paraules del Wall Street Journal i 200.000 paraules de la col·lecció TDT-4, respectivament. OntoNotes ha estat construït a partir del Penn Treebank per l'anotació sintàctica i del Penn PropBank per les estructures argumentals dels predicats. L'anotacio semàntica inclou entitats amb nom, sentits de les paraules (lligats a una ontologia) i relacions de coreferència. És distribuït pel Linguistic Data Consortium.

KNACK-2002 (HOSTE i DE PAUW, 2006) Un corpus de l'holandès que conté 267 documents extrets del setmanari flamenc Knack. Estan anotats manualment amb coreferència damunt de les etiquetes anotades semiautomàticament de PoS, *chunks* sintagmàtics i entitats amb nom.

TüBa-D/Z (HINRICHS *et al.*, 2005) Un banc d'arbres sintàctics de l'alemany basat en dades extretes del diari "die tageszeitung" (taz). Actualment comprèn 794.000 paraules anotades manualment amb informació semàntica i coreferencial. Per raons de drets d'autor dels textos originals, cal comprar un DVD del taz per aconseguir una llicència del corpus.

LiveMemories (RODRÍGUEZ *et al.*, 2010) Un corpus de l'italià en construcció que inclourà textos de la Viquipèdia en italià, blogs, notícies i diàlegs (MapTask). S'estan anotant automàticament amb sintaxi i manualment amb informació sobre coreferència (seguint l'esquema d'anotació ARRAU), concordança i entitats amb nom.

D'acord amb la lingüística funcional (HALLIDAY i HASAN, 1976; GUNDEL *et al.*, 1993), vaig adoptar un enfocament basat en la representació del discurs, que situa el fenomen de la coreferència dins del model de discurs projectat pels usuaris de la llengua i que substitueix la noció de "referents del món" per "referents del discurs" (KARTTUNEN, 1976; WEBBER, 1979; KAMP, 1981). El model de discurs es construeix dinàmicament i s'actualitza constantment, incloent no només les entitats explícitament mencionades, sinó també les que es poden inferir. En el camp computacional, l'opinió que les entitats pertanyen al món real ha predominat (NG, 2009; FINKEL i MANNING, 2008). Així, és només una minoria la que ha optat per la hipòtesi d'un model de discurs (POESIO, 2004*a*; BENGTSON i ROTH, 2008; DENIS i BALDRIDGE, 2009). Basar-se en el món real pot semblar que evita argot teòric innecessari, però desencadena una sèrie de problemes conceptuals, començant per les entitats imaginàries i hipotètiques: A quina entitat del món real es refereix *Superman*? O a quina entitat real ens referim quan un fa plans sobre el seu pròxim cotxe? En resum, en la definició de coreferència que donen VAN DEEMTER i KIBBLE (2000),

> SN_1 i SN_2 *corefereixen* si i només si Referent(SN_1) = Referent(SN_2),
> on Referent(SN) és una forma abreujada de "l'entitat referida per SN"

considero Referent (SN) com "l'entitat discursiva a què es refereix el SN en el model de discurs".

Dit això, el principi rector de la meva investigació era aconseguir un bon compromís entre la precisió lingüística i les possibilitats computacionals fent, per exemple, que les definicions fossin el més operatives possible. Per tal d'acotar l'àmbit d'estudi, calia posar algunes limitacions. Aquesta tesi se centra en la coreferència intradocumental,[8] incloent l'anàfora d'identitat de referència però excloent fenòmens que són estrictament anafòrics (HIRST, 1981) com l'anàfora d'identitat de sentit (3), l'el·lipsi (4), l'anàfora lligada (5) i l'anàfora pont (6), que implica que el lector faci una inferència mitjançant una relació diferent d'identitat per identificar l'àncora textual.

(3) La Lila conduïa [un cotxe] i la Maria també en conduïa [un].

(4) Al Jordi li van comprar una enorme caixa de [bombons] però en quedaven ja [pocs] [ø] al final del dia.

(5) [Cada cadena de televisió] va informar dels [seus] guanys.

(6) Vaig mirar dins de [l'habitació]. [El sostre] era molt alt.

De la mateixa manera, les relacions atributives (7), predicatives (8) i apositives (9) queden fora de l'àmbit del present estudi ja que no són relacions referencials. Així, segueixo esquemes d'anotació com MATE i OntoNotes que distingeixen entre coreferència i relacions predicatives (POESIO, 2004*b*; PRADHAN *et al.*, 2007*b*).

(7) Durant els dies que van seguir, van brollar deus de [foc] d'una dotzena d'orificis del volcà.

(8) El volcà Eyjafjallajökull és [un dels més grans d'Islàndia].

(9) El volcà Eyjafjallajökull, [un dels més grans d'Islàndia], havia estat en repòs durant gairebé dos segles.

Una segona limitació va ser concentrar-se en els actes de referència realitzats per SNs. Per SN em refereixo a pronoms, noms propis i sintagmes amb un nom comú com a nucli, que corresponen amb els "pronominals", "noms" i "nominals", respectivament, de l'ACE. Els termes "pronoms" i "SNs lèxics" s'utilitzen per distingir els dos últims grups del primer. Les expressions no nominals, però, no van ser del tot excloses, ja que l'anotació d'AnCora també inclou verbs, clàusules i segments discursius (*dixi discursiva* segons WEBBER [1979]). Vegeu-ne un estudi preliminar a RECASENS (2008).

Des del punt de vista computacional, volia fer recerca en tècniques d'aprenentatge automàtic ja que tenen un gran potencial per descobrir patrons i tendències

[8] Donada la seva funció discursiva, genuïnament la coreferència es dóna dins d'una unitat de discurs o al llarg d'una col·lecció de documents si tracten del mateix tema. Aquest treball, doncs, considera l'anomenada *coreferència entre documents* com una aplicació del PLN que pressuposa l'existència d'un discurs global subjacent que permet que diversos documents es tractin com un sol macrodocument.

generals que passen desapercebudes a l'ull humà, i des de mitjans dels anys noranta la resolució de la coreferència havia estat un objectiu de l'aprenentatge automàtic. Vaig fer ús de l'aprenentatge basat en memòria utilitzant TiMBL v.6.1.0 (DAELEMANS i BOSCH, 2005) després de provar altres paquets d'aprenentatge automàtic com màxima entropia (BERGER *et al.*, 1996) i arbres de decisió (QUINLAN, 1993). La preferència per TiMBL es va basar principalment en la robustesa davant d'espais de trets dispersos i les ajudes del paquet per a l'usuari, per exemple, el desglossament del rànquing dels trets d'aprenentage.

L'avaluació es va basar en una doble estratègia: (i) l'ús de les mesures més àmpliament acceptades (la mètrica MUC, B^3 i CEAF) per avaluar *quantitativament* el funcionament del sistema, i (ii) l'anàlisi manual d'errors d'una mostra de textos anotats automàticament per avaluar la sortida del sistema també *qualitativament*.

1.5.2 Indicis de ruptura

Al llarg d'aquesta investigació, vaig acumular una sèrie de troballes que suggerien l'existència de problemes de fons en la tasca de coreferència. Així, en un canvi de rumb, vaig desviar l'atenció de la qüestió de refinar la fase de resolució a la qüestió de reconsiderar el problema de la coreferència des del principi per tal de formular una solució viable. Aquí destaco els principals indicis que són mostra d'aquesta ruptura, sense entrar en els detalls tècnics que es poden trobar als articles dels capítols que segueixen.

Graus de referencialitat Anotar un corpus ens força a considerar cadascuna de les relacions, una a una, observada a les dades en comptes de seleccionar només les relacions que són fàcils i evidents. En aquest estudi, com que es pretenia anotar les relacions de coreferència entre SNs (capítol 2), aviat va sorgir la necessitat de distingir no només els sintagmes atributius (7), predicatius (8) i apositius (9) sinó també de separar tots els SNs referencials dels no referencials (10).

(10) El volcà Eyjafjallajökull, un dels més grans d'Islàndia, havia estat en repòs durant [gairebé dos segles] fins que va cobrar [vida] el capvespre del dia 20 de març de 2010, perceptible a[l principi] per l'emergència d'un núvol vermell que brillava per sobre de la vasta glacera que el cobreix. Durant els dies que van seguir, van brollar deus de foc d'una dotzena d'orificis del volcà, que arribaven fins a[ls 100 metres].

Si bé és possible identificar certes classes de no referencialitat com les expressions de durada (per exemple, *gairebé dos segles*), expressions de mesura (per exemple, *els 100 metres*) i frases fetes (per exemple, *al principi*), la distinció entre referencial i no referencial queda difuminada als extrems. Com a exemple, considereu *vida* a (10). Es troba a la frontera de gramaticalitzar-se i d'aquí la pèrdua del determinant com ha passat amb *anar a casa*. Aquests casos recolzen la idea de FRAURUD (1992) que ser un referent discursiu no és una qüestió de sí o no. Permetre graus

d'individuació, és a dir, diferents nivells de representació en el model de discurs, sembla concordar amb les dades observades.

Rellevància de les entitats unàries Atès que la coreferència és una relació binària, els projectes d'anotació solen marcar només les entitats de múltiples mencions. Això no és problemàtic si la referencialitat també es marca; altrament, però, sí que ho és, ja que llavors tots els SNs no coreferents es compten com a entitats unàries per defecte (ja que no hi ha altra manera d'extreure automàticament les entitats unàries a partir de l'anotació manual) i això provoca òbviament que un bon nombre de SNs que no són referencials quedin inclosos dins del conjunt de mencions. Tal dificultat es va trobar en voler comparar OntoNotes i AnCora (capítol 4) així com en extreure els conjunts de dades per a la tasca del SemEval dels corpus OntoNotes, KNACK-2002 i TüBa-D/Z (capítol 6).

En perspectiva, la qüestió de les entitats unàries té més implicacions de les inicialment esperades. Detectar la no coreferència és tan important com detectar la coreferència i és per això que alguns sistemes de coreferència detecten l'anomenada (encara que mal anomenada) "anaforicitat" per evitar tractar una menció com a subsegüent quan no ho és (NG, 2004; LUO, 2007; DENIS i BALDRIDGE, 2007). Les mencions que es classifiquen com a no anafòriques, però, segueixen considerant-se una vegada rera l'altra com a candidates de ser primeres mencions durant el procés de resolució. De fet, la preponderància de mencions aïllades (60% de tots els SNs, taula 4.1; 53% de totes les mencions, taula 2.3) deixa entreveure que un classificador automàtic capaç de detectar les entitats unàries seria de gran ajuda ja que permetria filtrar les mencions que no cal considerar ni com a subsegüents ni com a primeres. Dels experiments preliminars que vaig realitzar en aquesta direcció, però, no se'n desprèn que les mencions aïllades presentin cap propietat lingüística distintiva.

Pel que fa a l'avaluació, el gran nombre d'entitats unàries causa que el *baseline* de totes-unàries (és a dir, un sistema la sortida del qual conté una entitat per a cada menció) obtingui resultats tan alts (capítols 4 i 6), sobretot per als corpus sense restriccions de tipus d'entitats, com AnCora, OntoNotes, TüBa-D/Z i LiveMemories. La diferència en el nombre d'entitats unàries entre OntoNotes i ACE és la raó per la qual els sistemes de resolució puntuen més alt en el primer que en el segon d'acord amb les mesures basades en classes com B^3 i CEAF. STOYANOV *et al.* (2009) arriben a una conclusió similar pel que fa al motiu pel qual els sistemes avaluats sobre l'ACE obtenen millors resultats que els sistemes avaluats sobre el MUC. La principal font de problemes de les mètriques actuals és justament el tractament de les entitats unàries.

Distribució de les entitats L'atenció que s'ha dedicat al paper clau que tenen les cadenes de coreferència en la cohesió i la coherència del discurs ha conduït a donar per suposat que un bon nombre d'entitats discursives es mencionen més d'una vegada. Ja he argumentat en contra d'això amb relació a les entitats unàries, però

cal fer una altra observació. La majoria de les entitats de múltiples mencions es mencionen no moltes vegades, només *algunes* vegades. Es desprèn de les dades anotades que el tamany mitjà per entitat és d'entre tres i quatre mencions; de fet, el tamany més freqüent és de dues mencions (taula 2.5). En resum, el panorama general és que la distribució de les entitats d'un discurs està esbiaixada: una majoria d'entitats són unàries amb un paper perifèric i un segon gran grup d'entitats es mencionen un parell de vegades, la qual cosa deixa el nombre d'entitats que es mencionen més de dues vegades en aproximadament dues per document.

Això posa en relleu la divisió de KARTTUNEN (1976) entre referents de curta durada i de llarga durada, o la distinció de GROSZ i SIDNER (1986) entre entitats de focus local i de focus global. Hi ha hagut molts pocs intents (la mesura de "densitat de decadència" de DAUMÉ III i MARCU [(2005)] n'és un) de fer tal distinció en els sistemes de coreferència, però probablement resultaria útil. Perquè un sistema sigui capaç de decidir sobre la centralitat d'una entitat, cal implementar una estratègia que vagi més enllà de trets d'aprenentatge a nivell de parell.

Límits de la coreferència Una observació clau és que no totes les classes d'entitat tenen el mateix potencial de coreferir. Les persones i organitzacions (generalment introduïdes per noms propis) tendeixen a ser més coreferides que els llocs o les dates, per exemple. És una prova a favor de la hipòtesi que el tipus ontològic de les entitats diferencia nivells d'individuació (FRAURUD, 1992) i, per tant, de coreferència. És precisament entre els tipus més individualitzats, és a dir, persones i organitzacions, que es donen la majoria de desacords entre anotadors. Exemples com (11) i (12) no es poden classificar ni com a coreferents ni com a no coreferents.

(11) Aquí, durant segles, [els habitants] han tingut una relació gairebé mística amb Popo, arribant a creure que el volcà és un déu. Aquesta nit, [ø] temen que prendrà venjança.

(12) Per Aznar, la Monarquia Parlamentària "no solament és l'expressió de [l'Espanya moderna], sinó que a més ø és símbol d'estabilitat i permanència" ... Segons Aznar, la Corona pot "garantir i expressar" que [Espanya] pot tenir "més ambicions, propòsits i objectius."

Aquestes relacions dubtoses no són un fenomen marginal. Van ser conflictives no només per anotar el corpus Ancora, sinó també per entrenar el sistema que proposo sobre els corpus ACE i OntoNotes. La naturalesa problemàtica d'aquests exemples es va fer del tot evident en comparar els textos que estan anotats tant a ACE (13-a) com a OntoNotes (13-b) (entre claudàtors les mencions anotades com a coreferents).

(13) a. Ahir a la nit, a Tel Aviv, [jueus] van atacar un restaurant que té palestins contractats. "ø Volem la guerra", cantava [la multitud].

b. Ahir a la nit, a Tel Aviv, jueus van atacar un restaurant que té palestins contractats. "[ø] Volem la guerra", cantava [la multitud].

La diferent anotació de cada corpus revela –com també ho fa el baix o *tot just* acceptable grau d'acord entre anotadors aconseguit en projectes anteriors (POESIO i VIEIRA, 1998; MÜLLER, 2007; POESIO i ARTSTEIN, 2005)– un punt feble de l'actual definició de coreferència. La definició vigent és massa general per donar compte de tota la casuística observada en exemples reals. No fa esment de fenòmens metonímics, que abunden en la llengua, ni d'entitats infraespecificades, que jo he tractat breument en relació amb els objectes abstractes (RECASENS, 2008). El panorama resultant ens mostra la coreferència com un fenomen molt més complicat.

Absència de regles universals Poc a poc vaig anar assimilant la complexitat de la coreferència, a la vegada que duia a terme experiments amb aprenentatge automàtic. Aquests també van ajudar a posar al descobert les mancances de la tasca de resolució de la coreferència així com les limitacions de l'aprenentatge automàtic per resoldre el problema.

L'èxit que l'aprenentatge automàtic ha assolit en tasques de PLN com l'etiquetatge morfològic o l'anàlisi sintàctica es deu, en gran part, al fet que es poden aprendre a partir de propietats superficials (per exemple, els mots de l'entorn: *n*-grames) o la distribució d'etiquetes i paraules en certs contextos rellevants. Seguint un camí similar, la resolució de la coreferència ha tractat d'explotar marques morfològiques, sintàctiques i semàntiques, però no ha aconseguit resultats tan satisfactoris com els de l'etiquetatge morfològic o l'anàlisi sintàctica. Aquest fracàs es pot atribuir a tres causes principals (capítol 3). En primer lloc, sembla haver-hi ben poques regles que les relacions de coreferència compleixin sistemàticament (HOSTE, 2005; DENIS, 2007). Això també queda demostrat per l'estancament en què s'arriba a mesura que s'augmenta el tamany de les dades d'entrenament (capítol 4). En segon lloc, els trets d'aprenentatge que s'han utilitzat fins al moment no aconsegueixen recollir aspectes pragmàtics que són imprescindibles per detectar relacions de coreferència úniques i específiques. I encara se segueixen descobrint efectes nous (ARNOLD i GRIFFIN, 2007). Per últim, s'observen complexes interaccions entre els trets que no controlem (LUO i ZITOUNI, 2005), la qual cosa explica la *limitada* millora dels resultats aconseguida amb els 351 trets de URYUPINA (2008) en comparació amb els 12 trets de SOON *et al.* (2001). Això també explica la importància que té fer una selecció de trets (NG i CARDIE, 2002*b*; HOSTE, 2005). En resum, tots aquests factors impedeixen que els algorismes d'aprenentatge automàtic puguin desenvolupar un model de coreferència suficientment general i universal.

Estudis psicolingüístics i cognitius en anglès (ARNOLD *et al.*, 2000; GORDON *et al.*, 1993; CRAWLEY *et al.*, 1990; STEVENSON *et al.*, 1994; KEHLER *et al.*, 2008), en castellà (CARREIRAS i GERNSBACHER, 1992) i en català (MAYOL i CLARK, 2010) han presentat proves empíriques per a molts dels trets utilitzats pels sistemes de resolució de la coreferència dels darrers quinze anys. En la mesura que els experiments han demostrat que aquests trets fan que els lectors prefereixin

interpretar els pronoms d'una manera i no d'una altra, s'ha suposat que els mateixos trets haurien de funcionar igualment bé per als sistemes automàtics, ja sigui en forma de restriccions i preferències o en forma de trets d'aprenentatge. No obstant això, una raó important per la qual no han tingut els resultats esperats és el salt que hi ha entre els estudis de laboratori o amb una orientació lingüística i les dades reals, on tots els fenòmens interactuen i es donen a la vegada. Això no és negar el valor i la importància d'aquests estudis, però cal avaluar adequadament la seva contribució a la tasca en qüestió. Com KRAHMER (2010) assenyala, els psicolingüistes i els lingüistes computacionals tenen objectius diferents. Els primers s'interessen a mostrar efectes globals i a aprendre sobre la memòria humana; els segons, en canvi, s'interessen a obtenir bons resultats i, per tant, es preocupen per cadascun dels punts de les dades que el sistema no processa correctament.

Teories més generals de la referència com la Teoria de l'accessibilitat (ARIEL, 1988), la Jerarquia dels referents donats (GUNDEL *et al.*, 1993) o la Teoria del *centering* (GROSZ *et al.*, 1995) estableixen criteris segons els quals les expressions referencials se situen al llarg d'una escala. Tanmateix, els múltiples factors que intervenen en la valoració del grau d'"accessibilitat" o l'"estatus cognitiu" expliquen per què aquestes nocions són tan difícils de representar computacionalment i desemboquen en problemes d'implementació (POESIO *et al.*, 2004*a*). La distància, per exemple, és un factor crucial per determinar el grau d'accessibilitat, però no és l'únic (ARIEL, 2001). En aquest sentit, TETREAULT (1999) esmenta usos inconsistents del gènere, l'estudi basat en corpus de BARBU *et al.* (2002) troba que gairebé una quarta part dels pronoms plurals corefereixen amb un constituent que no és plural i POESIO *et al.* (2004*a*) assenyalen que la coherència que es manté entre les entitats d'enunciats diferents és molt menys forta d'allò inicialment esperat.

Els exemples següents il·lustren alguns dels problemes que els sistemes de resolució de la coreferència haurien de ser (però no són) capaços de resoldre. Els exemples d'ús real inclouen nombrosos contraexemples a trets com la concordança de nombre (14), la definitud com a marcador d'accessibilitat (15), la indefinitud com a marcador d'informació nova (16) i fins i tot al tret de mateix-nucli (17) (18).

(14) a. [Madeleine Albright] es reuneix demà amb [Ehud Barak] i [el president palestí Yasser Arafat]. S'espera que [ø] es reuneixin a la tarda.
b. Madeleine Albright es reuneix demà amb [Ehud Barak] i [el president palestí Yasser Arafat]. S'espera que [ø] es reuneixin per separat amb Albright.

(15) a. [El volcà Eyjafjallajökull, un dels més grans d'Islàndia,] havia estat en repòs durant gairebé dos segles ... van brollar deus de foc d'una dotzena d'orificis d[el volcà], arribant fins als 100 metres.
b. [El volcà Eyjafjallajökull, un dels més grans d'Islàndia,] havia estat en repòs durant gairebé dos segles.

(16) a. [Un nou estudi que detalla la càrrega de treball no compensada dels metges de família] assenyala la necessitat de canviar la forma en què

són remunerats.

b. Possiblement [Postville] s'està posant al dia amb la resta d'Amèrica ... Les fàbriques van ajudar a impulsar el desenvolupament econòmic en [una ciutat que portava molt temps estancada].

(17) a. Gillian Finley, ABC, comença a [la Gaza palestina]. Avui, a [Gaza], soldats israelians han obert foc contra col·legials que llançaven pedres.

b. El president Clinton era a [Irlanda del Nord] quan es va assabentar de la decisió de la Cort Suprema ... Clinton va agrair al govern d'[Irlanda] haver acceptat dos presoners.

(18) a. [Un centenar d'artistes] participarà a l'acte ... D[els nombrosos artistes que s'han ofert a participar en aquest homenatge], la meitat ho farà al principi i l'altra meitat al final de la celebració.

b. [Un centenar d'artistes] participarà a l'acte ... No ha estat possible comptar amb [tots els nombrosos artistes que s'han ofert a participar en aquest homenatge] per qüestió de temps.

Pragmàtica Atès que els algorismes d'aprenentatge no disposen d'informació explícita sobre la pragmàtica o el coneixement del món, els resulta molt complicat discriminar entre (14-a) i (14-b). De la mateixa manera, en el cas dels SNs definits (la forma de SN més freqüent en català i castellà), tot i que se sol considerar que es refereixen a una entitat prèviament introduïda (15-a), més del 50% de les vegades són la forma de mencions aïllades o primeres mencions (15-b) (FRAURUD, 1990; POESIO i VIEIRA, 1998; RECASENS *et al.*, 2009*a*; RECASENS, 2009). D'altra banda, se sol considerar que els SNs indefinits compleixen la funció contrària, és a dir, esmenten una entitat desconeguda per primera vegada (16-a), però de nou, aquesta regla no està exempta d'excepcions (16-b). Pel que fa als noms propis, tant poden introduir un SN com coreferir amb una entitat prèviament introduïda (taula 2.3). De fet, totes les teories que situen les expressions referencials en una escala (ARIEL, 1988; GUNDEL *et al.*, 1993) coincideixen en què factors pragmàtics addicionals poden prevaler sobre els principis proposats. Per afegir un exemple més, HERVÁS i FINLAYSON (2010) mostren que el 18% de les expressions referencials de textos periodístics i narratius són descriptives, és a dir, proporcionen informació addicional no necessària per distingir el referent. Això pot ser contrari al principi que identifica les descripcions definides llargues amb marcadors de baixa accessibilitat (ARIEL, 1988).

En la tònica de trets no generalitzables, el de mateix-nucli sembla ser el més robust. De tots ells, és clarament el tret que resol el major nombre de relacions amb el menor nombre d'errors, encara que tampoc funciona sempre (VIEIRA i POESIO, 2000; SOON *et al.*, 2001; URYUPINA, 2008; ELSNER i CHARNIAK, 2010), com s'exemplifica amb els casos positius (17-a) (18-a) enfront dels negatius (17-b) (18-b). Tot i això, la resolució de relacions de coreferència en què hi participen noms propis o SNs lèxics que no tenen el mateix nucli es manté com

un dels problemes més difícils (HAGHIGHI i KLEIN, 2010). Millorar la cobertura dels sistemes amb el mínim cost de precisió és només factible fins a un 80% de mesura F. El 20% restant és realment perjudicial per a la qualitat lingüística dels resultats, encara que és una qüestió que ha rebut poca atenció. A aquest punt hi tornaré més endavant. Sens dubte, els trets que s'han proposat són tots importants en algun aspecte, però s'escapa als mètodes actuals la manera en què fonts de coneixement i trets diferents col·laboren i interactuen en conjunció.

Efectes del preprocessament Una altra expectativa que no es va complir plenament està relacionada amb l'ús d'informació de preprocessament automàtic en contraposició amb informació morfològica i sintàctica acurada. Per tal de determinar la mesura en què els resultats empitjoren quan no es disposa d'informació d'un estàndard de referència, el mateix sistema de coreferència s'ha aplicat als mateixos textos variant la font d'informació de preprocessament (capítol 4). Sorprenentment, el descens no va ser gaire pronunciat, la qual cosa pot ser deguda a dues raons. En primer lloc, trets rellevants com el de mateix-nucli no es veuen afectats per la qualitat del preprocessament. En segon lloc, l'algorisme d'aprenentatge reordena els trets de tal manera que la informació morfològica guanya posicions respecte a la informació sintàctica, la qual cosa minimitza el soroll de les eines automàtiques. Amb un bon conjunt de trets bàsics (superficials) doncs, l'aprenentatge pot funcionar gairebé igual de bé sense necessitat de trets rics que depenguin d'una anàlisi sintàctica profunda.

Les conseqüències d'utilitzar preprocessament automàtic són més greus pel que fa a la detecció dels límits de mencions (STOYANOV *et al.*, 2009), la qual cosa URYUPINA (2008) identifica com la causa principal d'enllaços coreferencials erronis. Utilitzar límits de mencions predits automàticament provoca un descens dels resultats tal com mostra la taula 1.3 si comparem les puntuacions dels sistemes que han estat avaluats amb ambdós tipus de mencions: reals i del sistema (NICOLAE i NICOLAE, 2006; NG, 2008; HAGHIGHI i KLEIN, 2009; RAHMAN i NG, 2009). Aquest va ser un error de disseny que vam cometre a l'hora de definir la configuració de l'avaluació *gold* i la regular en la tasca SemEval (capítol 6): el fet de donar els límits de les mencions reals només en el *gold* ens va privar de poder treure conclusions sobre les millores dels resultats gràcies a la utilització de preprocessament acurat. A diferència de STOYANOV *et al.* (2009), que defensen que la configuració experimental amb límits de mencions reals és "poc realista" ja que simplifica considerablement la tasca de coreferència, considero que el problema de detectar les mencions és d'ordre diferent. És un problema que pertany, en primer lloc, a la sintaxi i, en segon lloc, a la detecció de la referencialitat. Per tant, avaluar les tasques de resolució de la coreferència i de detecció de mencions com una sola tasca no només es presta a la confusió sinó que fa que resultats diferents no siguin comparables.

Els *baselines* de mateix-nucli i totes-unàries En els últims anys, no s'ha parat de suggerir nous trets i models de resolució més sofisticats, però sorprèn comparar els resultats d'aquests models recents amb els obtinguts per dos *baselines* ben simples (capítol 4): (i) lligar totes les mencions que comparteixen el mateix nucli ("*baseline* de mateix-nucli") i (ii) no classificar cap menció com a coreferent sinó totes com a unàries ("*baseline* de totes-unàries"). La superioritat de mateix-nucli és corroborada pel fet que es troba a la base de la mesura predictiva de resultats de coreferència de STOYANOV *et al.* (2009), que, donat un conjunt de dades, en prediu la complexitat de resolució. A més, MARKERT i NISSIM (2005) identifiquen el gran nombre de SNs definits que estan governats pel simple tret de mateix-nucli com un dels problemes per comparar els resultats dels algorismes de coreferència.

Allò més preocupant és que aquests *baselines*, sobretot el de totes-unàries, no s'acostumen a incloure en les publicacions sobre resolució de la coreferència. CARDIE i WAGSTAFF (1999) sí que donen els resultats de mateix-nucli amb el corpus de test de MUC-6 i admeten que "té un rendiment millor del que s'esperaria". SOON *et al.* (2001) i URYUPINA (2007) també donen aquest *baseline* amb el corpus de test de MUC-7. La diferència de 5 punts percentuals entre els seus *baselines* respectius s'explica pel fet que utilitzen mencions del sistema. Avaluen, per tant, sobre conjunts de mencions que no són iguals. El sistema de SOON *et al.* (2001) supera mateix-nucli només per un 5%, mentre que el de URYUPINA (2007) sobrepassa el seu *baseline* per un 15%.

Segons el corpus, el *baseline* de totes-unàries pot aconseguir tot sol resultats de fins el 84% B^3 i 73% CEAF (per a OntoNotes). El mateix *baseline* minva a un 67% B^3 i 50% CEAF per a ACE a causa del menor nombre d'entitats unàries. Quan s'afegeix mateix-nucli, els resultats per a ACE pugen fins a un 76% B^3 i un 66% CEAF (cal tenir en compte el gran nombre de noms propis del corpus ACE). Per contra, lligar massa mencions com a coreferents té com a resultat una puntuació de la mètrica MUC major per a ACE (68%) que per a OntoNotes (56%). La importància d'aquestes xifres resideix en què els sistemes actuals com el de LUO *et al.* (2004) obtenen un 77% B^3, 73% CEAF i 81% MUC (per al corpus ACE-2), resultats que no estan lluny dels dos *baselines* gens artificiosos, sobretot si es té en compte l'enorme esforç invertit en aquests sistemes.

Resultats buits d'informació Deficiències en la manera com es calculen les mesures d'avaluació provoquen que senzills *baselines* com el de totes-unàries o el contrari, la sobreunió de mencions, obtinguin resultats massa alts. Amb la mètrica MUC, per exemple, agrupar una menció dins l'entitat equivocada se sanciona el doble que agrupar dues entitats reals (POESIO *et al.*, en preparació), d'aquí el seu biaix cap a la sobreunió. No només la mètrica MUC sinó també B^3 i CEAF estan esbiaixades cap a diferents tipus de sortida (capítols 4 i 5). Per això, la puntuació final d'un sistema acaba depenent més de les característiques del corpus que de l'estratègia de resolució. La classificació de MUC situarà en primer lloc el sistema la sortida del qual tingui el major nombre d'enllaços i B^3 premiarà sistemes amb

una alta cobertura que aconsegueixin detectar entitats unàries i unir les mencions que tenen el mateix nucli –encara que uneixin mencions no coreferents (ELSNER i CHARNIAK, 2010). CEAF, d'altra banda, tot i que aconsegueix un millor equilibri, encara està fortament influït pel gran nombre d'entitats unàries. El desacord entre les tres mesures a l'hora d'ordenar els sistemes les fa poc útils per al seu propòsit, és a dir, per avaluar models de resolució de la coreferència. Així doncs, no va ser possible treure conclusions definitives sobre la tasca compartida del SemEval (capítol 6), ja que cada mesura classifica els sistemes participants en un ordre diferent (taula 6.5).

El problema de la qualitat empitjora a mesura que s'estén la pràctica de donar *només* els resultats numèrics, sense una mostra real de la sortida del sistema. Com a mínim, SOON *et al.* (2001) i URYUPINA (2008) realitzen una anàlisi d'errors. L'experiència revela que mirant a ull el resultat, podem adonar-nos que mencions com *el nou president* i *l'antic president* estan lligades per pura coincidència de nucli. Tal error és encara molt comú en els sistemes d'última generació (HAGHIGHI i KLEIN, 2010). La qualitat dels sistemes de coreferència actuals està lluny de ser satisfactòria i les mesures d'avaluació de què disposem, en lloc de ser una eina útil en aquest sentit, contribueixen a amagar els resultats. Naturalment una anàlisi qualitativa pressuposa una comprensió de la definició de la tasca, la qual cosa ens porta de nou a alguns dels punts plantejats més amunt. A falta d'aquesta comprensió, la cursa cap al desenvolupament de sistemes corre el risc de convertir-se en una cursa cap a la sintonització dels sistemes amb el corpus de test i invalida la noció de l'existència d'un estàndard de referència real i independent. Aquí, de fet, és on hem començat: els corpus anotats, on la comprensió del fenomen queda reflectida, són determinants de la cadena de fets subsegüents.

1.5.3 Noves línies

Després d'haver observat el seguit d'inconvenients apuntats a 1.5.2, vaig veure clar que perquè la tasca de coreferència tingués sentit i, sobretot, fos útil per a la comunitat de PLN, calia reconsiderar diversos aspectes fonamentals. Suposava un repte que anava més enllà de l'abast inicial d'aquesta investigació, però de manera natural vaig fer els primers (tres) passos en aquesta direcció, que es presenten en aquest apartat. Els dos primers passos es van produir durant el camí, com a resultat directe de la recerca en curs; el tercer era més ambiciós però necessari per completar aquesta tesi.

CISTELL El sistema basat en entitats de LUO *et al.* (2004) obria una nova porta a la resolució de la coreferència, però continua sent "una àrea que necessita més investigació", com ells mateixos assenyalen. L'enfocament que he escollit s'afegeix al conjunt de models basats en entitats ja que crea un sistema, CISTELL, que manipula les entitats discursives com si fossin cistells en creixement (capítol 4). La noció d'un cistell en creixement és semblant a la *targeta arxivadora* de HEIM (1983) dins la semàntica d'intercanvi d'arxius, on hi ha una fitxa per a cada entitat

Menció: *els seus companys de professió*	
Atribut	**Valor**
És un pronom	fals
Nucli	companys
És-un	col·lega, company de feina
Hiperònim	associat
Gènere	masculí
Nombre	plural
Determinant	els | seus
Numeral	—
Tipus d'entitat amb nom	—
Complements	de professió
Nucli de l'oració	ser
Funció sintàctica	adjunt
Posició en paraules	17
Posició en mencions	6
Posició en oracions	2

Taula 1.4: Informació continguda en un "cistell"

discursiva de manera que la informació de les referències posteriors s'hi pot anar emmagatzemant a mesura que progressa el discurs.

Al principi del discurs, a cada menció se li assigna un cistell que conté atributs de la menció com ara el nucli, el tipus, els complements, etc., de la manera exemplificada a la taula 1.4. Part de la informació s'extreu directament del text i l'altra part de recursos externs com WordNet. La característica més rellevant dels cistells és que poden *créixer* ingerint altres cistells i incorporant els seus atributs. Un esbós d'aquest procés es mostra a la figura 1.1, on es veu un document a la fase intermèdia del seu processament, amb els cistells representats simbòlicament. Quan dos cistells es classifiquen com a coreferents, són agrupats immediatament en un cistell que encara pot seguir creixent.

El procés general de resolució s'inspira en POPESCU-BELIS *et al.* (1998). La figura 1.2 és una il·lustració congelada del procés de creixement i mostra els diferents moviments que CISTELL pot fer en un moment donat. El cistell rosat és el que s'està examinant. Tant pot ser engolit i contribuir al creixement del cistell gran de l'esquerra (figura 1.2(a)), com pot ser engolit per un dels cistells més petits (i en creixement) de la dreta (figura 1.2(b)), com pot quedar-se unari, amb l'oportunitat de créixer més endavant en el discurs (figura 1.2(c)).

El quid de la qüestió és el procés de creixement, és a dir, decidir si dos cistells s'han o no de fusionar per créixer. La decisió és una o altra en funció de si els dos cistells són unaris o si un d'ells ja ha començat a créixer. Les decisions del primer tipus són directes ja que només es basen en la probabilitat de coreferència donada pel classificador de parells, mentre que les del segon tipus es poden prendre de diverses maneres tenint en compte els diferents parells formats amb cadascun dels

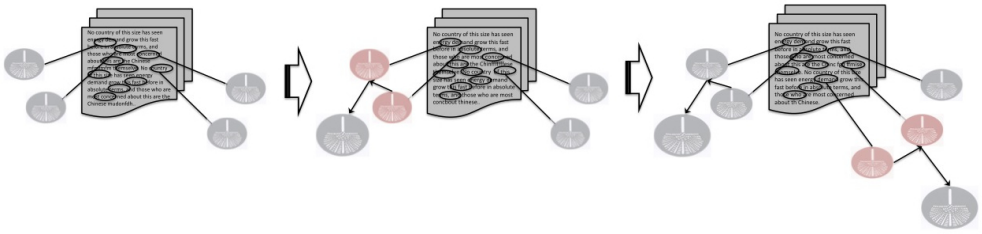

Figura 1.1: Representació del procés de resolució a CISTELL

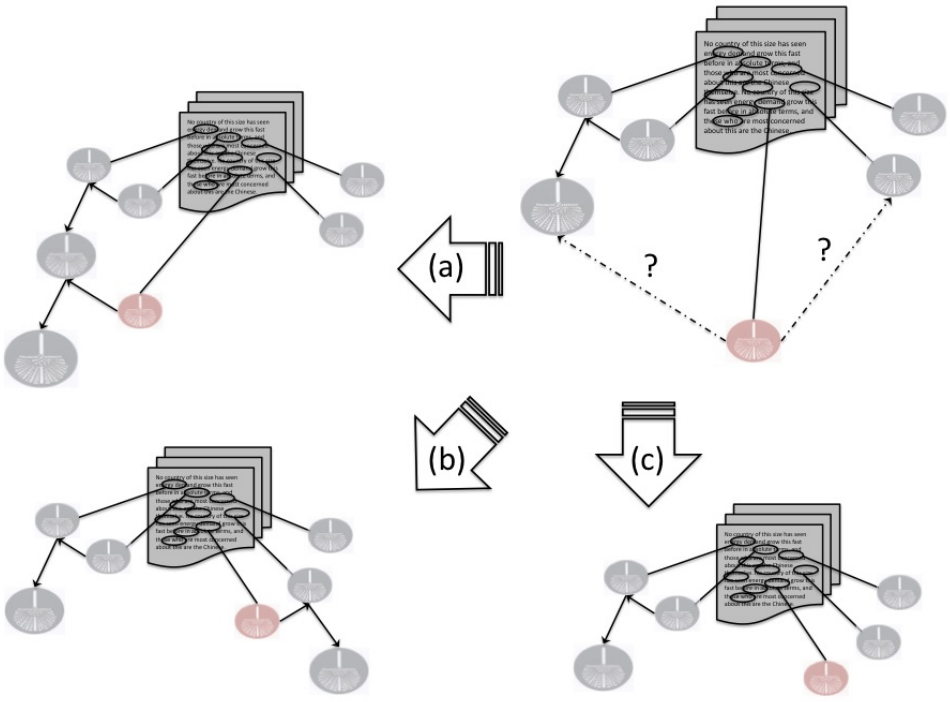

Figura 1.2: Representació del procés de creixement de cistells a CISTELL. (a) Creixement cap a l'esquerra, (b) Creixement cap a la dreta, (c) Cistell unari.

cistells dins del cistell en creixement. A aquestes decisions les anomeno *encaix*. Es defineix un paràmetre que especifica el nombre d'encaixos necessaris perquè un cistell sigui engolit per una entitat en creixement. Dels diferents valors que es van provar (qualsevol encaix, qualssevol dos encaixos, ..., tot encaixos), els millors resultats s'obtenen quan s'obliga que tots els parells siguin coreferents. A aquest requisit m'hi refereixo com a *encaix fort*. A l'extrem oposat, permetre que un únic parell compatible sigui garantia suficient perquè es produeixi la fusió (és a dir, *encaix feble*) es tradueix en una sobreunió.

A diferència de LUO *et al.* (2004), per tant, els resultats millors obtinguts per l'estratègia d'encaix fort (enfront de l'encaix feble) són una prova dels efectes beneficiosos d'utilitzar una estratègia més global. Mantenir constància de la història de cada entitat discursiva és util per capturar la major quantitat d'informació sobre una entitat que ens proporciona el text. D'altra banda, però, les estratègies globals poden tenir efectes adversos si no es dissenyen degudament. Els trets de distància, per exemple, només tenen sentit si es té en compte el concepte d'antecedent (KEHLER, 1997). Per tant, aquests trets funcionen per als pronoms i els SNs lèxics, però poden introduir soroll en el cas dels noms propis, que són més sensibles a trets com la freqüència de l'entitat (HAGHIGHI i KLEIN, 2010).

En principi, els cistells són il·limitats i es poden enriquir amb tants atributs com es cregui convenient i estiguin disponibles –amb informació de "dins" del text i amb coneixement contextual i del món de "fora" del text– la qual cosa permet codificar molts dels trets necessaris per resoldre les relacions de coreferència. El classificador de coreferència s'entrena conjuntament per resoldre la coreferència i per detectar les entitats noves en el discurs. Això s'aconsegueix mitjançant la generació d'instàncies negatives d'entrenament que, a diferència de SOON *et al.* (2001), inclouen no només mencions coreferents sinó també mencions d'entitats unàries. Encara que no va ser factible explotar al màxim CISTELL a causa dels obstacles que es van trobar durant el camí (apartat 1.5.2), és un sistema que ofereix un marc ple de possibilitats per donar cabuda al gir teòric que presento més avall (vegeu "La coreferència com a contínuum").

BLANC Per superar les mancances detectades de les mesures MUC, B^3 i CEAF, vaig partir de l'índex de Rand (RAND, 1971) per idear BLANC (de les sigles en anglès, *BiLateral Assessment of Noun-phrase Coreference*, és a dir, 'avaluació bilateral de la coreferència entre sintagmes nominals'), una nova mesura que té en compte no només els enllaços de coreferència sinó també els de no coreferència i, sobretot, que els dóna la mateixa importància (capítol 5). D'aquesta manera, les entitats unàries ni s'ignoren ni reben més pes que les entitats de múltiples mencions independentment de la freqüència d'ocurrència. La propietat interessant d'aplicar Rand per a la coreferència és que la suma conjunta de tots els enllaços de coreferència i de no coreferència és constant per a un determinat conjunt de n mencions.

En la configuració per defecte de BLANC, els dos tipus d'enllaç (coreferent i no coreferent) compten igual per a la puntuació final, la qual cosa incrementa el

	MUC			B^3			CEAF	BLANC		
	P	C	F	P	C	F	P / C / F	P	C	blanc
AnCora - Castellà										
1,	–	–	–	100	73,32	84,61	73,32	49,21	50,00	49,60
2,	55,03	37,72	44,76	91,12	79,88	**85,13**	75,96	77,63	58,57	62,90
3,	48,22	44,24	46,14	86,21	80,66	83,34	76,30	74,08	59,54	63,53
4,	45,64	51,88	48,56	80,13	82,28	81,19	75,79	69,14	66,80	**67,89**
5,	45,68	36,47	40,56	86,10	79,09	82,45	**77,20**	69,82	62,69	65,43
6,	43,10	35,59	38,98	85,24	79,67	82,36	75,23	69,05	62,79	65,27
7,	45,73	65,16	**53,75**	68,50	87,71	76,93	69,21	55,80	79,52	58,15
OntoNotes - Anglès										
1,	–	–	–	100	72,68	84,18	72,68	49,24	50,00	49,62
2,	55,14	39,08	45,74	90,65	80,87	**85,48**	76,05	77,36	62,64	67,19
3,	47,10	53,05	49,90	82,28	83,13	82,70	75,15	73,32	66,92	69,59
4,	47,94	55,42	51,41	81,13	84,30	82,68	78,03	71,53	70,36	**70,93**
5,	48,27	47,55	47,90	84,00	82,27	83,13	78,24	70,67	66,39	68,27
6,	50,97	46,66	48,72	86,19	82,70	84,41	**78,44**	74,82	67,87	70,75
7,	47,46	66,72	**55,47**	70,36	88,05	78,22	71,21	55,73	77,42	58,17

Taula 1.5: Resultats de la mètrica BLANC per a la taula 4.3, capítol 4.
1. = TOTES UNÀRIES; 2. = MATEIX NUCLI; 3. = MATEIX NUCLI + PRONOMS; 4. = ENCAIX FORT; 5. = ENCAIX SUPERFORT; 6. = ENCAIX ÒPTIM; 7. = ENCAIX FEBLE

pes de les entitats de múltiples mencions. No obstant això, BLANC es defineix com una mesura ponderada amb un paràmetre α que permet a l'usuari decidir si s'ha de donar més pes als enllaços coreferents o als no coreferents. Dividir així per la meitat la recompensa entre els enllaços de coreferència i els de no coreferència implica que cada tipus de lligam tingui un límit màxim de cobertura del 50%. Això penalitza directament els sistemes la sortida dels quals conté un nombre gran d'entitats unàries i massa pocs enllaços de coreferència o, en el cas extrem, els sistemes de "coreferència" que no fan altra cosa que retornar entitats unàries. Dos dels desideràtums principals darrera de la definició de BLANC són (i) no fer cas omís de cap dels dos tipus d'enllaç i (ii) aconseguir l'equilibri més satisfactori entre ells. La mesura ha estat provada en diferents corpus: ACE (amb el seu conjunt restringit de tipus semàntics), OntoNotes i Ancora.

Com que BLANC és una nova mesura, no la vaig incloure als experiments amb CISTELL que descriu el capítol 4, però val la pena fer-ho aquí. Les taules 1.5, 1.6 i 1.7 reprodueixen les taules 4.3, 4.5 i 4.7, però inclouen els resultats de la mètrica BLANC al costat dels de MUC, B^3 i CEAF. Se'n desprèn que BLANC es distingeix clarament de B^3 i CEAF pel que fa al *baseline* de totes-unàries (primera fila). A causa de l'absència de l'impuls inicial donat per la identificació d'entitats unàries, la resta de les puntuacions de BLANC sempre estan per sota de B^3 i en general per sota o lleugerament per sobre de CEAF. A l'altre extrem, i a diferència de MUC, BLANC també castiga considerablement la sobreunió (l'ENCAIX FEBLE, setena fila, tendeix a incloure molts enllaços, tant de correctes com d'erronis).

Curiosament, B^3 i CEAF no coincideixen mai en el sistema més ben qualificat, mentre que BLANC i CEAF estan d'acord dues de les sis vegades. Als

	MUC			B^3			CEAF	BLANC		
	P	C	F	P	C	F	P / C / F	P	C	blanc
Esquema d'OntoNotes										
1,	–	–	–	100	72,68	84,18	72,68	49,24	50,00	49,62
2,	55,14	39,08	45,74	90,65	80,87	**85,48**	76,05	77,36	62,64	67,19
3,	47,10	53,05	49,90	82,28	83,13	82,70	75,15	73,32	66,92	**69,59**
4,	46,81	53,34	49,86	80,47	83,54	81,97	76,78	68,80	69,19	68,99
5,	46,51	40,56	43,33	84,95	80,16	82,48	76,70	66,36	62,01	63,83
6,	52,47	47,40	49,80	86,10	82,80	84,42	**77,87**	71,80	67,60	69,46
7,	47,91	64,64	**55,03**	71,73	87,46	78,82	71,74	55,30	76,13	57,45
Esquema d'ACE										
1,	–	–	–	100	50,96	67,51	50,96	47,29	50,00	48,61
2,	82,35	39,00	52,93	95,27	64,05	**76,60**	66,46	88,80	60,99	66,35
3,	70,11	53,90	60,94	86,49	68,20	76,27	68,44	81,56	64,71	69,66
4,	64,21	64,21	64,21	76,92	73,54	75,19	**70,01**	75,51	71,07	**73,04**
5,	60,51	56,55	58,46	76,71	69,19	72,76	66,87	72,34	65,77	68,38
6,	67,50	56,69	61,62	82,18	71,67	76,57	69,88	76,94	69,00	72,17
7,	63,52	80,50	**71,01**	59,76	86,36	70,64	64,21	62,74	83,19	66,45

Taula 1.6: Resultats de la mètrica BLANC per a la taula 4.5, capítol 4.
1. = TOTES UNÀRIES; 2. = MATEIX NUCLI; 3. = MATEIX NUCLI + PRONOMS; 4. = ENCAIX FORT; 5. = ENCAIX SUPERFORT; 6. = ENCAIX ÒPTIM; 7. = ENCAIX FEBLE

	MUC			B^3			CEAF	BLANC		
	P	C	F	P	C	F	P / C / F	P	C	blanc
Esquema d'OntoNotes										
1,	–	–	–	100	72,66	84,16	72,66	49,07	50,00	49,53
2,	56,76	35,80	43,90	92,18	80,52	**85,95**	76,33	79,95	62,34	67,32
3,	47,44	54,36	50,66	82,08	83,61	82,84	74,83	73,44	68,30	70,53
4,	52,66	58,14	55,27	83,11	85,05	84,07	**78,30**	73,86	74,74	**74,29**
5,	51,67	46,78	49,11	85,74	82,07	83,86	77,67	71,27	67,51	69,20
6,	54,38	51,70	53,01	86,00	83,60	84,78	78,15	74,31	70,96	72,50
7,	49,78	64,58	**56,22**	75,63	87,79	81,26	74,62	60,00	78,89	64,22
Esquema d'ACE										
1,	–	–	–	100	50,42	67,04	50,42	47,32	50,00	48,62
2,	81,25	39,24	52,92	94,73	63,82	**76,26**	65,97	87,43	61,09	66,36
3,	69,76	53,28	60,42	86,39	67,73	75,93	**68,05**	81,05	64,50	69,37
4,	58,85	58,92	58,89	73,36	70,35	71,82	66,30	72,08	67,69	**69,60**
5,	56,19	50,66	53,28	75,54	66,47	70,72	63,96	70,68	63,56	66,23
6,	63,38	49,74	55,74	80,97	68,11	73,99	65,97	73,36	65,24	68,29
7,	60,22	78,48	**68,15**	55,17	84,86	66,87	59,08	60,02	80,08	62,27

Taula 1.7: Resultats de la mètrica BLANC per a la taula 4.7, capítol 4.
1. = TOTES UNÀRIES; 2. = MATEIX NUCLI; 3. = MATEIX NUCLI + PRONOMS; 4. = ENCAIX FORT; 5. = ENCAIX SUPERFORT; 6. = ENCAIX ÒPTIM; 7. = ENCAIX FEBLE

diagrames de dispersió de la figura 1.3 hi distingim el corpus com una variable rellevant per trobar correlacions significatives entre les mesures, especialment en el cas de BLANC. Hi ha una correlació positiva entre BLANC i CEAF tant a ACE ($\tau = 0{,}82$)[9] com a OntoNotes ($\tau = 0{,}57$), entre BLANC i MUC a ACE ($\tau = 0{,}53$) i a OntoNotes ($\tau = 0{,}48$), i en menor grau entre BLANC i B^3 però només a ACE ($\tau = 0{,}43$). CEAF està també correlacionat positivament a ACE amb B^3 ($\tau = 0{,}62$) i MUC ($\tau = 0{,}46$). La correlació entre BLANC i les altres mesures és menor a OntoNotes ja que aquest corpus té un major nombre d'entitats unàries. Com ja he comentat, la característica distintiva de BLANC és el seu tractament de les entitats unàries. És, per tant, una mesura preferible a les altres per comparar resultats de coreferència sobre corpus que estan anotats amb el conjunt complet de mencions.

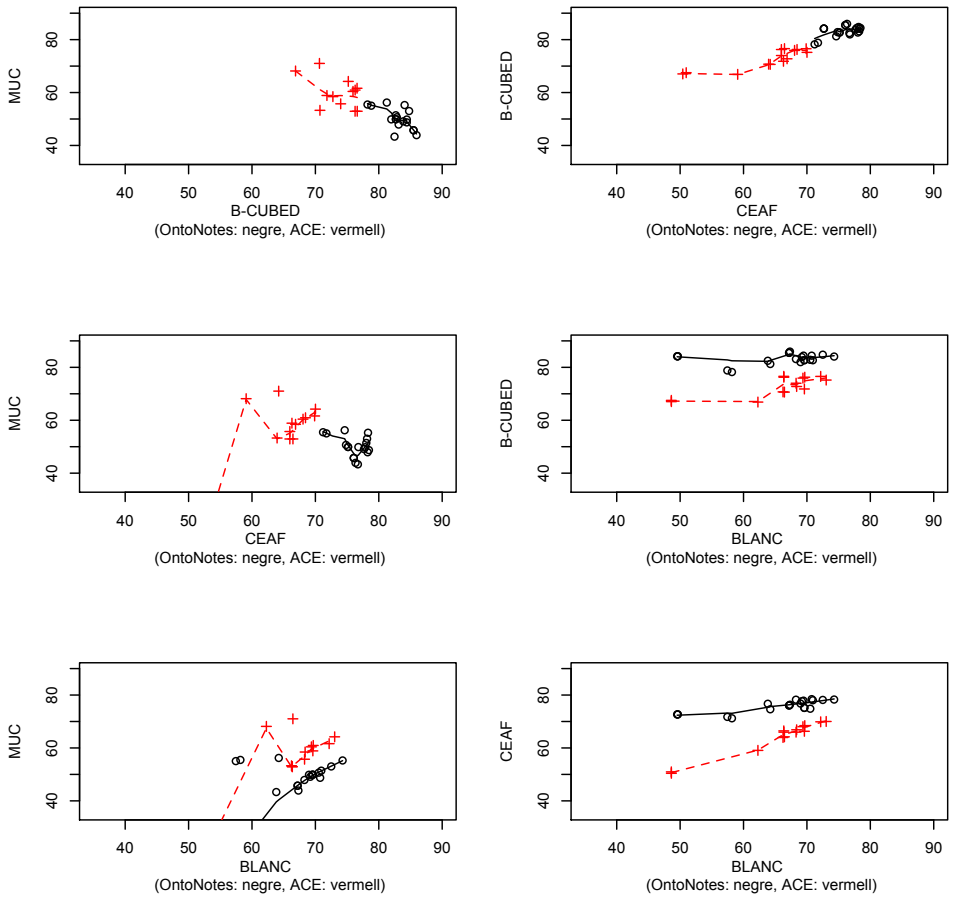

Figura 1.3: Diagrames de dispersió dels parells de MUC, B^3, CEAF i BLANC

[9] Totes les correlacions es mesuren amb la tau de Kendall (τ) i són significatives si p-valor $< 0{,}01$.

L'anàlisi quantitativa és recolzada per una avaluació qualitativa. A l'apèndix A hi figura una mostra de les sortides de sistemes per als mateixos dos documents d'OntoNotes. Per al primer document (nbc_0030, apèndix A.1), les quatre mesures situen l'ENCAIX FORT en primer lloc, però discrepen en el rànquing de les tres altres sortides de CISTELL. A B^3 i CEAF, a diferència de MUC, clarament no els agrada l'ENCAIX FEBLE a causa del seu biaix cap a les entitats unàries. BLANC puntua les sortides que no són ENCAIX FORT de manera molt similar i en un interval molt inferior, la qual cosa evidencia la seva inferior qualitat.

Per al segon document (voa_0207, apèndix A.2), la sortida guanyadora, per ser la menys sorollosa, és l'ENCAIX SUPERFORT i és també la sortida millor puntuada per les quatre mesures. Una vegada més, MUC és l'únic que puntua alt l'ENCAIX FEBLE. Com a prova de la major potència discriminatòria de BLANC, l'ENCAIX FORT i l'ENCAIX ÒPTIM reben la mateixa puntuació en termes de CEAF, mentre que BLANC mostra una lleugera preferència pel primer model que està en sintonia amb una anàlisi qualitativa de les dues sortides.

La mesura BLANC ha estat utilitzada públicament per primera vegada en el concurs SemEval, d'acord amb els objectius de la tasca (capítol 6). A més de les sortides de CISTELL, l'apèndix A.1 també inclou els resultats dels sis sistemes participants. Això possibilita la comparació dels resultats numèrics presentats a la taula 6.5 amb els textos anotats automàticament. Des d'una perspectiva qualitativa, la millor sortida és la del sistema CORRY-C, que és també el millor classificat segons CEAF i BLANC. En canvi, B^3 situa RELAXCOR en primer lloc, la sortida del qual inclou un nombre molt reduït d'enllaços de coreferència, i MUC premia SUCRE, que tendeix a produir massa enllaços. Són proves addicionals del biaix de B^3 cap a les entitats unàries així com del biaix de MUC cap a la sobreunió, respectivament. La majoria de les mètriques situen TANL-1 i UBIU en últim lloc i això concorda amb la mala qualitat de les seves sortides. Finalment, no és possible comparar BART amb les altres sortides al mateix nivell ja que només va participar en l'avaluació regular i, per tant, els resultats amb mencions reals no estan disponibles.

La coreferència com a contínuum Sorprèn que ben pocs estudis (VERSLEY, 2008; POESIO i ARTSTEIN, 2005; CHAROLLES i SCHNEDECKER, 1993) hagin considerat les implicacions teòriques i els problemes que comporta la definició actual de coreferència. Si ni tan sols els humans es posen d'acord en què és i què no és una relació de coreferència, no s'haurien d'imposar expectatives poc realistes de la meta on poden arribar els sistemes automàtics de resolució de la coreferència. Com ja s'ha assenyalat a l'apartat 1.5.2, les dades reals mostren un seguit de relacions –com s'exemplifica a (12) i (13)– que són inexplicables en termes dicotòmics de coreferència i no coreferència.

Trobo que la font del problema radica en la hipòtesi que la coreferència es basa en termes estrictes d'identitat excloent. En lloc d'això, introdueixo el concepte de *quasi-identitat* juntament amb un model continu de la coreferència (capítol 8).

Això deixa espai per a la comprensió i la representació dels casos en què la relació d'identitat no és total sinó parcial: X i Y són els mateixos pel que fa a certes característiques, però difereixen en almenys una característica. En el llenguatge quotidià, sovint diem frases de l'estil *Ell i jo som iguals* quan el significat implícit és que ell i jo tenim moltes coses en comú, o ocupem una posició social similar, encara que òbviament no som la mateixa persona i som diferents en molts aspectes.

Si ens traslladem a l'àmbit de la coreferència, suggereixo que els nostres judicis de coreferència estan directament relacionats amb el nivell de granularitat en què categoritzem les entitats d'un discurs donat. Aquest nivell de granularitat ve determinat per les intencions comunicatives i la coherència del discurs. Com a exemple, considereu la categorització d'*Espanya* en dos contextos diferents: una descripció històrica afavorirà una lectura no coreferencial d'*Espanya* i *l'Espanya moderna*, mentre que en un altre context, com per exemple una notícia del dia, probablement les dues expressions seran usades indistintament, afavorint-ne així una lectura coreferencial.

La postulació d'un contínuum de coreferència és coherent amb els diferents graus en què sembla funcionar la referencialitat (FRAURUD, 1992). De fet, com més referencial és una entitat, més es pot especificar. Si bé té sentit concebre Espanya ja sigui a nivell general i sincrònic ja sigui subdividida en *l'Espanya del segle XVII*, *l'Espanya moderna*, etc., seria estrany pensar de la forquilla amb què un menja en termes de *la forquilla d'ahir* o *la forquilla de demà*. El fet que no totes les entitats puguin ser concebudes al mateix nivell d'especificitat ens porta de nou al que he apuntat més amunt sobre el paper que tenen els tipus ontològics en la individuació d'una entitat. A l'igual que les categories s'organitzen al llarg d'un contínuum, la coreferència es produeix al llarg d'un contínuum des de la identitat a la no identitat, amb un interval de relacions de quasi-identitat entre aquests dos extrems.

D'acord amb BYBEE (2010), qui considera que l'estructura lingüística deriva d'aplicar processos cognitius de domini general (per exemple, la categorització o el *chunking*), suggereixo tres operacions cognitives de categorització que donen compte de les relacions de quasi-identitat que es mantenen entre entitats que comparteixen la majoria de valors dels atributs, però no tots. En funció de si una entitat discursiva augmenta, sobreescriu o anul·la un valor d'una entitat existent, distingeixo entre *especificació*, *reenfocament* i *neutralització*, respectivament. Les dues primeres operacions creen noves funcions indicials ja que mostren una faceta diferent d'una entitat complexa, mentre que l'última combina dues o més entitats similars en una sola categoria, neutralitzant així funcions indicials potencials. Aquesta triple distinció ens permet tenir un marc flexible que, entre altres avantatges, ofereix una justificació operativa per al tractament de la metonímia des de la perspectiva de la coreferència, la qual cosa encara és un tema controvertit i una de les causes principals de desacord entre anotadors.

L'especificació, el reenfocament i la neutralització es representen dins del marc de la teoria d'espais mentals de FAUCONNIER (1997). Es tracta d'un marc conceptualment apropiat i que és consistent amb les idees que fonamenten el sistema CIS-

TELL. Els espais mentals són estructures mentals abstractes que construïm mentre pensem i parlem a l'efecte de la comprensió local i sobre les quals es projecten les entitats discursives. Fauconnier reconeix que les eines de la lògica formal fallen quan s'enfronten amb la gamma de fenòmens presents al llenguatge natural. Els espais mentals, però no la lògica formal, poden explicar casos com el jo partit (19) i la coreferència partida (20) el significat dels quals requereix dividir un referent en dos.

(19) a. Si [jo] fos tu, [m']odiaria.
b. Si [jo] fos tu, m'odiaria [a mi mateix].

(20) Si [Woody Allen] hagués nascut bessons, [ø] haurien sentit llàstima mútua, però [ø] no hi va néixer, així que només sent llàstima d'ell mateix.

Els participants del discurs conceptualitzen les entitats amb una sèrie d'atributs associats que tenen valors específics en funció de l'espai concret. Quan s'introdueix una nova entitat que té propietats que o bé entren en conflicte amb una entitat discursiva ja existent o bé n'eliminen detalls, llavors cal construir un nou espai mental amb l'entitat quasi-idèntica corresponent. La coreferència partida és una de les relacions que requereixen un canvi d'espai mental. He identificat deu classes d'atributs que requereixen aquests canvis quan es canvien els seus valors i les he organitzades en una jerarquia (RECASENS *et al.*, 2010*a*). Aquesta jerarquia estableix les maneres més freqüents que desencadenen en relacions de quasi-identitat. Els detalls es troben cap al final de la tesi al capítol 8. L'apèndix B complementa el capítol amb el recull d'extractes de quasi-identitat que es van utilitzar en l'estudi d'acord entre anotadors lligat a la tipologia de quasi-identitat, i l'apèndix C inclou les classificacions dels anotadors de les relacions ressaltades en els extractes.

1.6 Aportacions principals

La principal contribució d'aquesta tesi és la visió crítica i alhora constructiva de diversos aspectes de la tasca de resolució de la coreferència, que van des de la resolució i l'avaluació a l'anotació de corpus i la teoria subjacent. El resultat final és el desenvolupament d'un nou enfocament de la coreferència basat en la síntesi de dades extretes de corpus, operacions cognitives de domini general i la teoria d'espais mentals de FAUCONNIER (1997). A diferència de l'enfocament dicotòmic actualment estès, el model continu de la coreferència que proposo deixa lloc a relacions intermèdies de quasi-identitat per poder explicar satisfactòriament les dades reals.

Les aportacions principals d'aquesta tesi són:

- Els corpus AnCora del català i el castellà (un total de gairebé 800.000 paraules) anotats amb informació sobre la coreferència, juntament amb un esquema d'anotació que resol deficiències d'enfocaments anteriors i que incorpora etiquetes específiques per al català i el castellà.[10]

[10] http://clic.ub.edu/corpus/ancora

- Una llista de més de quaranta-cinc trets d'aprenentatge que són examinats en un model de resolució de la coreferència basat en parells per al castellà. Els resultats manifesten que els trets són poc informatius individualment, però que participen en interaccions complexes i impredictibles.
- Un model de resolució de la coreferència basat en entitats anomenat CISTELL, que combina i integra les decisions de parells al nivell del discurs.
- Una anàlisi dels punts febles de les mesures actuals d'avaluació de la coreferència que enfosqueixen l'avaluació dels sistemes de coreferència i dificulten les comparacions directes entre els sistemes de l'estat de l'art.
- La implementació de l'índex de Rand per a l'avaluació de la coreferència en la nova mesura BLANC, que té en compte –per igual en la configuració per defecte– tant els enllaços de coreferència com els de no coreferència.
- Una comparació entre la coreferència i la paràfrasi, destacant les similituds i les diferències per tal d'apuntar possibles àrees de col·laboració mútua entre la resolució de la coreferència i l'extracció de paràfrasis.
- L'organització de la tasca compartida en resolució de la coreferència del SemEval així com la publicació a la web dels conjunts de dades, el mòdul avaluador i la documentació que es van crear per a la tasca.[11]
- Un petit corpus de 60 fragments de quasi-identitat, una tipologia de relacions de quasi-identitat i un estudi d'acord entre anotadors que en demostra l'estabilitat.
- Un model continu de la coreferència que va des de la identitat a la no identitat passant per relacions de quasi-identitat així com tres operacions cognitives de categorització (l'especificació, la neutralització i el reenfocament) que permeten explicar les diferents etapes al llarg d'aquest contínuum.
- La identificació d'una sèrie de febleses de l'enfocament actual de la resolució de la coreferència que apunten cap a la necessitat de reconsiderar diferents aspectes de la tasca.

Aquestes contribucions constitueixen el contingut dels capítols 2 al 8 que segueixen a continuació.

[11] http://stel.ub.edu/semeval2010-coref/

Part I

ANOTACIÓ DE CORPUS AMB COREFERÈNCIA

CAPÍTOL 2

AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan

Marta Recasens and M. Antònia Martí

University of Barcelona

Published in *Language Resources and Evaluation*, 44(4):315–345

Abstract This article describes the enrichment of the AnCora corpora of Spanish and Catalan (400k each) with coreference links between pronouns (including elliptical subjects and clitics), full noun phrases (including proper nouns), and discourse segments. The coding scheme distinguishes between identity links, predicative relations, and discourse deixis. Inter-annotator agreement on the link types is 85%-89% above chance, and we provide an analysis of the sources of disagreement. The resulting corpora make it possible to train and test learning-based algorithms for automatic coreference resolution, as well as to carry out bottom-up linguistic descriptions of coreference relations as they occur in real data.

Keywords Coreference · Anaphora · Corpus annotation · Annotation scheme · Reliability study

2.1 Introduction

Producing a text requires us to make multiple references to the entities the discourse is about. Correspondingly, for a proper understanding of the text, we have to identify the entity each linguistic unit refers to and link those that are **coreferent**, that is, those that stand in an *identity of reference* relation. Following Webber's

(1979) discourse model, coreference does not take place between real-world entities but between discourse entities, i.e., the (mental) entities in a listener's evolving model of the discourse, which may or may not correspond to something in the outside world.

Although often treated together with anaphora, coreference is different (VAN DEEMTER i KIBBLE, 2000). Coreference involves the semantico-referential level of language, since in order to identify those expressions (whether anaphoric or non-anaphoric) that refer to the same discourse entity, we must first understand their semantics and find their referents; while anaphora occurs at the textual level: in order to interpret an empty (or almost empty) textual element—an anaphor—like *el cicle* 'the cycle' in (1-a),[1] we need to go back in the text to find its antecedent (*el seu primer cicle de concerts* 'their first cycle of concerts'). Thus, anaphora and coreference work independently, although they can co-occur. We distinguish **anaphoric coreference** (1-a) from **definite coreference** (1-b), where the last expression (*Endemol, productora del programa Gran Hermano* 'Endemol, the production company for the Big Brother programme') is understood without the need of going back in the text. Finally, (1-c) shows that not all anaphoric relations are coreferent: *les de moros i cristians* 'those of Moors and Christians' is anaphoric, since the lexical head *festes* 'festivals' is retrieved from the previous expression *festes de lluita de classes* 'class struggle festivals,' but each expression refers to a different entity, i.e., they do not corefer.

(1) a. (Cat.) Els integrants del Cor Vivaldi assagen les peces d*el seu primer cicle de concerts*. En aquesta primera edició d*el cicle* ...
'The members of the Vivaldi Choir are rehearsing the compositions for *their first cycle of concerts*. In this first event of *the cycle* ...'

b. (Sp.) El director general de Telefónica Media, Eduardo Alonso, dijo hoy que la alianza con *la productora Endemol* ha beneficiado más a *la empresa holandesa* que a Telefónica. ... esta alianza ha beneficiado más a John de Mol y a los socios de *Endemol, productora del programa Gran Hermano.*
'The director-general of Telefónica Media, Eduardo Alonso, said today that the alliance with *the Endemol production company* has benefitted *the Dutch company* more than Telefónica. ... this alliance has been of more benefit to John de Mol and the partners of *Endemol, the production company for the Big Brother programme*.'

c. (Cat.) A algú se li acudirà organitzar *festes de lluita de classes*, igual que existeixen *les de moros i cristians*.
'Somebody will think of organizing *class struggle festivals*, just as there are *those of Moors and Christians*.'

The goal of anaphora resolution is to fill the empty (or almost empty) expressions

[1] All the examples throughout the article have been extracted from the AnCora-CO corpora. Those preceded by (Cat.) come from Catalan and those by (Sp.) from Spanish.

in a text, i.e., to find an antecedent for each anaphoric unit so that the latter is linked to the mention its interpretation depends on. Coreference resolution, on the other hand, aims to establish which (referential) noun phrases (NPs) in the text point to the same discourse entity, thus building coreference chains. Hence, while the outputs of anaphora resolution are antecedent-anaphor pairs, the outputs of coreference resolution are collections of mentions[2] of different types (referential pronouns and their antecedents, proper nouns, definite NPs, discourse segments, etc.) that refer to the same discourse entity. Solving coreference can imply solving anaphora, i.e., anaphoric coreference. This article presents a language resource that can be used for coreference resolution as well as for limited anaphora resolution.[3]

Given its cohesive nature, coreference is a key element in the comprehensive interpretation of a text and, by extension, an interesting object of study both in computational and theoretical linguistics. By building the coreference chains present in a text, we can identify all the information about one entity. From a computational perspective, the identification of coreference links is crucial for a number of applications such as information extraction, text summarization, question answering, and machine translation (MCCARTHY i LEHNERT, 1995; STEINBERGER *et al.*, 2007; MORTON, 1999). From a linguistic point of view, capturing the way a discourse entity is repeatedly referred to throughout a discourse makes it possible to obtain the different ways an entity can be linguistically expressed. Besides, empirical data on the way coreference relations are actually expressed provide a way to test hypotheses about the cognitive factors governing the use of referring expressions such as those suggested by ARIEL (1988) and GUNDEL *et al.* (1993).

The importance of the coreference resolution task in information extraction led to its inclusion in two Message Understanding Conferences (MUC)—1995 and 1998—and in the more recent ACE evaluation programs, as well as the Anaphora Resolution Exercise (ARE) (ORASAN *et al.*, 2008). It will also be one of the tasks at SemEval-2010 (RECASENS *et al.*, 2009*b*). Due to the complexity inherent in coreference, limitations of rule-based approaches (HOBBS, 1978; BALDWIN, 1997; LAPPIN i LEASS, 1994; MITKOV, 1998) may be overcome by machine learning techniques, which allow to automate the acquisition of knowledge from annotated corpora (SOON *et al.*, 2001; NG i CARDIE, 2002*b*; LUO *et al.*, 2004). The information extraction conception which is behind MUC and ACE is basically interested in finding all the information about a particular entity, thus conflating referential and predicative links, for example. Since this lack of precision in defining coreference (against predicative links and other related phenomena) is problematic, one of our goals was delimiting the boundaries of the concept of "coreference" to annotate a corpus in a systematic and coherent way.

[2]Following the terminology of the Automatic Content Extraction (ACE) program (DODDINGTON *et al.*, 2004), a *mention* is defined as an instance of reference to an object, and an *entity* is the collection of mentions referring to the same object in a document.

[3]To obtain anaphoric coreference pronouns from AnCora-CO, one just needs to extract the pronouns that are included in an entity. By convention, we can assume that their antecedent corresponds to the previous mention in the same entity.

This article describes the annotation of the Spanish and Catalan AnCora corpora (Section 2.2) with coreference information. Currently, AnCora-CO comprises two 400,000-word corpora annotated with coreference links (distinguishing identity from discourse deixis and predicative relations) between pronouns, full noun phrases (including proper nouns), and discourse segments. AnCora-CO makes it possible to train corpus-based coreference resolution systems for Spanish and Catalan, as well as to infer linguistic knowledge about the way coreference relations occur in real data. Three main assets make AnCora-CO a valuable language resource: its size, its target languages, and the quality of its annotation—the coding scheme is the result of a study that takes into account linguistic evidence and schemes previously proposed for English (Section 2.3). The following sections provide details about the coding scheme (Section 2.4), the annotation tool (Section 2.5), statistics on the tags (Section 2.6), and inter-annotator agreement (Section 2.7). The article concludes with a discussion of the results (Section 2.8).

2.2 The corpora

Corpora annotated with coreference information are scarce. Those most widely used have been developed for English within the MUC and ACE evaluation programs (HIRSCHMAN i CHINCHOR, 1997; DODDINGTON *et al.*, 2004). However, both datasets call for improvement from a linguistic perspective: the former has been criticized for the underlying theoretical implications of the coding guidelines (VAN DEEMTER i KIBBLE, 2000), whereas the latter restricts coreference to relations between seven specific entity types.[4] Other domain-specific corpora have also been or are being developed for English within ongoing annotation tasks (MITKOV *et al.*, 2000; POESIO, 2004*a*; HOVY *et al.*, 2006; POESIO i ARTSTEIN, 2008).

Coreferentially annotated corpora are even scarcer for languages other than English. Among these few we find Czech, German and Dutch (KUČOVÁ i HAJIČOVÁ, 2004; HINRICHS *et al.*, 2004; STEDE, 2004; HOSTE, 2005). For Spanish, there is the coreferentially annotated corpus developed for ACE-2007,[5] but again the coreference links annotated are limited to the set of ACE-like entity types. There are also two small corpora of Spanish oral narratives and dialogues (BLACKWELL, 2003; TABOADA, 2008), but they are highly restricted to pronominal references for the purpose of studying the neo-Gricean maxims and centering theory, respectively.

The annotation of coreference in AnCora constitutes an additional layer added on top of existing in-line annotations (TAULÉ *et al.*, 2008): morphological (POS and lemmas), syntactic (constituents and functions) and semantic (argument structures, thematic roles, semantic verb classes, NEs, and WordNet nominal senses). The AnCora-CO corpus is split into two datasets: the Spanish corpus (AnCora-CO-Es), and the Catalan corpus (AnCora-CO-Ca). Each consists of 400,000 words derived

[4] ACE-2004 entity types include: person, organization, geo-political entity, location, facility, vehicle and weapon.

[5] http://projects.ldc.upenn.edu/ace/docs/Spanish-Entities-Guidelines_v1.6.pdf

from newspaper and newswire articles: 200,000 words from the Spanish and Catalan versions of *El Periódico* newspaper, and 200,000 words from the EFE newswire agency[6] in the Spanish corpus, and from the ACN newswire agency[7] in the Catalan corpus. AnCora-CO is the largest multilayer annotated corpus of Spanish and Catalan. It is freely available from http://clic.ub.edu/corpus/en/ancora.[8]

2.3 Linguistic issues

Given that coreference is a pragmatic linguistic phenomenon highly dependent on the situational context, it does not fall under the topics traditionally dealt with by descriptive Spanish or Catalan grammars apart from some occasional references (BOSQUE i DEMONTE, 1999; SOLÀ, 2002). When analysing real data, we come across a wide range of units (e.g., pronouns in quoted speech) and relations (e.g., metonymic relations) which cannot easily be identified as coreferent or otherwise. Besides, although there are theoretical linguistic studies for English, coreference shows certain language-specific patterns. For instance, Spanish and Catalan make extensive use of elliptical pronouns in subject position, whereas English uses overt pronouns and shows a different distribution of definite NPs.

This endeavour at annotation met two needs—that of delimiting the boundaries of the concept of "identity of reference," and the need to deal with specific aspects of Spanish and Catalan. The design of the annotation scheme for AnCora-CO began by considering corpus data and listing problematic issues which the scheme needed to address specifically. Our approach was to develop a coding scheme with sufficient criteria to decide which tags had to be used and for what; that is, a scheme from which the corpora could be consistently annotated. Following is a discussion of key issues concerning coreference annotation—illustrated with real data from the two languages—providing an overview of the coreference annotation in AnCora-CO by explaining how each of them was dealt with in the actual annotation.

1. *Elliptical pronouns*. Spanish and Catalan are pro-drop languages that allow pronominal subjects to be omitted if no contrast is being made. Coreference relations can thus involve elliptical elements.[9]

[6] http://www.efe.es

[7] http://www.acn.cat

[8] At present, a total of 300,000 words for each AnCora-CO corpus are freely downloadable from the Web. An additional subset of 100,000 words is being kept for test purposes in future evaluation programs.

[9] Elliptical subject pronouns are marked with ø and with the corresponding pronoun in brackets in the English translation.

(2) (Cat.) La mitjana d'edat d*els ramaders* és de 47 anys i ø tenen una jornada laboral de 73 hores setmanals.
'The average age of *the stock farmers* is 47 years and *(they)* have a 73-hour working week.'

Since elliptical subjects were inserted when AnCora was syntactically annotated (they have their own NP node), it is easy to include them when coding a coreference link. Elliptical subjects that are pleonastic –which are not as frequent as they are in English– are not annotated, as in the Catalan pattern ø *és que...* '*It* is that...'

2. *Clitic pronouns*. Object personal pronouns appear as clitic forms in the two languages under consideration. Postverbal clitics take a different form in each language: Spanish clitics are adjoined to the verbal head (3-a), while the clitic is joined with a hyphen in Catalan (3-b).

(3) a. (Sp.) La intención es reconocer *el gran prestigio que tiene la maratón* y unir*lo* con esta gran carrera.
'The aim is to recognize *the great prestige that the Marathon has* and join|*it* with this great race.'
b. (Cat.) ø va demanar un esforç per *assimilar l'euro amb rapidesa* i no deixar-*ho* per més endavant.
'(She/He) called for an effort *to assimilate the euro quickly* and not postpone-*it* for later.'

Clitic pronouns are generally referential, except for inherent clitics that form a single unit of meaning with the verb (e.g., Sp. *jugársela*, Cat. *jugar-se-la* 'to risk it'). For spelling reasons, incorporated clitics do not have their own token in AnCora-Es. Hence, the verbal node is annotated for coreference,[10] while Catalan clitics have their own NP node.

3. *Quoted speech*. Deictic first and second person pronouns (4-a) become anaphoric in quoted speech, and can be thus linked to the corresponding speaker. The first person plural pronoun presents two atypical uses that need to be taken into account. The royal *we* (4-b), which is used when somebody speaks not in his/her own name, but as the leader of a nation or institution, is linked to such an organization, if this appears explicitly in the text. Similarly, the editorial *we* (4-c) is commonly used in newspaper articles when referring to a generic person as *we*, as if the writer is speaking on behalf of a larger group of citizens. Since there is no explicit group to which these pronouns can be linked, first mentions are considered to have no antecedent,

[10]Two guiding principles in the morphological annotation of AnCora were (a) to preserve the original text intact, and (b) to assign standard categories to tokens, so that a category such as "verb-pronoun" for verbs with incorporated clitics was ruled out.

and subsequent mentions are linked with the closest previous editorial *we* pronoun.

(4) a. (Sp.) *El guardameta del Atlético de Madrid, A. Jiménez*, cumplió ayer uno de sus sueños al vencer al Barcelona. "*ø* Nunca había ganado al Barcelona".
'*The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona. "*(I)* had never beaten Barcelona".'

b. (Cat.) En paraules d'un dels directius de *l'agència*, "Ramón y Cajal *ens* va deixar tirats".
'In the words of one of the *agency*'s board members, "Ramón y Cajal left *us* in the lurch".'

c. (Cat.) L'efecte 2000 era un problema real, encara que *tots* hem ajudat a magnificar-lo.
'The 2000 effect was a real problem, even though *we all* helped to magnify it.'

4. *Possessives*. Possessive determiners and possessive pronouns might have two coreferential links: one for the thing(s) possessed (5-a) and one for the possessor (5-b). The former is marked at the NP level, whereas the latter is marked at the POS level.[11]

(5) a. (Cat.) La diversitat pel que fa a la nacionalitat d*els músics d'Il Gran Teatro Amaro* és un dels factors importants, tot i que *els seus components* sempre han mostrat interès.
'The diversity of nationality among *the musicians of Il Gran Teatro Amaro* is one of the important factors, although *its members* have always shown interest.'

b. (Cat.) La diversitat pel que fa a la nacionalitat dels músics d'*Il Gran Teatro Amaro* és un dels factors importants, tot i que els *seus* components sempre han mostrat interès.
'The diversity of nationality among the musicians of *Il Gran Teatro Amaro* is one of the important factors, although *its* members have always shown interest.'

5. *Embedded NPs*. Coreference often involves NPs embedded within a larger NP. For instance, between the NPs *el presidente de los Estados Unidos* 'the president of the U.S.' and *el presidente del país* 'the president of the country,' two links are encoded: one between the entire NPs, and one between *los Estados Unidos* 'the U.S.' and *el país* 'the country.' However, if an embedded NP functions as an apposition, then the maximal NP principle applies,

[11] Possessive determiners are not considered NPs according to the syntactic annotation scheme.

by which only the largest stretch of NP is to be annotated. For this reason, a phrase such as *la ciudad de Los Angeles* 'the city of Los Angeles' is considered to be atomic.
The maximal NP rule also applies to constructions of the type "the members of (the set)." In *los jugadores de Argentina* 'the players of Argentina,' *Argentina* refers to the football team[12] rather than the country, and, since the team is equivalent to the players, coreference is marked for the entire NP.

6. *Split antecedent*. Plural NPs can refer to two or more individuals mentioned separately in the text.

   (6) a. (Sp.) ø Propongo abrir la campaña con *un debate político general* y cerrarla con *otro*, aunque Ríos advirtió que él está dispuesto a que en *esos debates* participen los cabezas de otros partidos.
   '(I) intend to start the campaign with *a general political debate* and end|it with *another one*, although Ríos indicated that he is prepared to allow the heads of other parties to participate in *those debates*.'
   b. (Cat.) Un partit obert fins al final per les ocasions de gol a *les dues porteries* ... El Racing va buscar *la porteria contrària*.
   'A game open until the end due to the goal-scoring chances at *both ends* . . . Racing plugged away at *the opposing goalmouth*.'

   Cases like (6-a) are resolved by building an entity resulting from the addition of two or more entities: entity1+entity2... The converse (6-b), however, is not annotated: mentions that are subentities of a previous entity are not linked, since this implies a link type other than coreference, namely part-of or set-member.

7. *Referential versus attributive NPs*. Not all NPs are referential, they can also be attributive. Schemes such as MUC and ACE treat appositive (7-a) and predicative (7-b) phrases as coreferential. Regarding MUC, VAN DEEMTER i KIBBLE (2000) criticize it for conflating "elements of genuine coreference with elements of anaphora and predication in unclear and sometimes contradictory ways." Besides, if attributive NPs are taken as coreferential, then other predicate-like NPs such as the object complement of the verb *consider* should be too (7-c), and might easily result in incorrect annotations.

   (7) a. (Cat.) El grup de teatre *Proscenium*.
   'The theatrical company *Proscenium*.'

[12] The fact that *Argentina* is marked as NE-organization provides a clue for the annotators to apply the maximal NP principle. This principle, however, turned out to be a source of inter-annotator disagreement (see Section 2.7.2).

b. (Cat.) L'agrupament d'explotacions lleteres és *l'únic camí*.
'The unification of dairy operations is *the only way*.'
c. (Sp.) El Daily Telegraph considera a Shearer *"el hombre del partido"*.
'The Daily Telegraph considers Shearer *"the man of the match"*.'

To be loyal to the linguistic distinction between referential and attributive NPs, nominal predicates and appositional phrases are not treated as coreference in AnCora-CO. However, given that NPs identifying an entity by its properties can be useful for automatic coreference resolution, such relations are kept under the "predicative link" tag (see Section 2.4.2), which parallels the division between identical and appositive types followed in the OntoNotes annotation (PRADHAN *et al.*, 2007*b*). Keeping referential and attributive links apart makes it possible to use AnCora-CO at the user's discretion: either under a fine-grained definition of coreference or under a coarse one, obliterating the distinction between the two links in the latter case.

8. *Generic versus specific NPs*. Coreference links can occur on a specific or a more generic level. We decided that these two levels should not be mixed in the same coreference chain since the referential level is not the same. This is especially relevant for time-dependent entities, since a generic celebration (e.g., *the Olympic Games*) differs from specific instantiations (e.g., *the Barcelona Olympic Games*). Likewise, a function type (e.g., *the unemployment rate*) takes different values according to time and place (e.g., *the lowest unemployment rate in Spain at 6.6%*). Thus, these NPs are not annotated as coreferent.

9. *Metonymy*. The referent referred to by a word can vary when that word is used within a discourse, as echoed by Kripke's (1977) distinction between "semantic reference" and "speaker's reference." Consequently, metonymy[13] can license coreference relations between words with different semantic references (8).

(8) (Sp.) *Rusia* llegó a la conclusión ... *Moscú* proclamó ...
'*Russia* came to the conclusion ... *Moscow* proclaimed ...'

Metonymy within the same newspaper article is annotated as a case of identity, since, despite the rhetorical device, both mentions pragmatically corefer. It is just a matter of how the entity is codified in the text. The transitivity test (see Section 2.4.2 below) helps annotators ensure that the identity of reference is not partial but complete.

[13] Metonymy is the use of a word for an entity which is associated with the entity originally denoted by the word, e.g., *dish* for *the food on the dish*.

10. *Discourse deixis*. Some NPs corefer with a previous discourse segment (9).[14] Since linking NPs with non-NP antecedents adds complexity to the task, and not all coreference resolution systems might be able to handle such relations, discourse deixis is kept separate as a different link type (see Section 2.4.2).

    (9) (Sp.) *Un pirata informático consiguió robar los datos de 485.000 tarjetas de crédito ...* El robo fue descubierto...
    '*A hacker managed to steal data from 485,000 credit cards ...* The theft was uncovered ...'

11. *Bound anaphora*. Although this relation has been treated as coreference in annotation schemes such as MUC, it expresses a relation other than coreference and therefore is not annotated in AnCora-CO. If in (10-a) *cada una* 'each' was taken as coreferent, then by the transitivity test[15] it would follow that *se quedaron con dos EDF y Mitsubishi* 'EDF and Mitsubishi took two,' a total of two licenses—not four—were bought.
    In contrast, coreference is allowed in (10-b) since, by being distributed into each of the components, *cada equipo* 'each team' results in a whole that equals the sum of the parts.

    (10) a. (Sp.) EDF y Mitsubishi participaron en la licitación de licencias para construir centrales eléctricas y se quedaron con dos *cada una*.
    'EDF and Mitsubishi participated in the bidding for licenses to build power stations and took two *each*.'
    b. (Sp.) *Brasil* buscará el pase a la final ante *los vigentes campeones, los australianos*. Los números uno de *cada equipo*, Rafter y Kuerten, abrirán el fuego en la primera jornada.
    '*Brasil* will be looking to pass to the final against *the current champions, the Australians*. The number ones of *each team*, Rafter and Kuerten, will open the first day's play.'

12. *Bridging reference*. Bridging relations (CLARK, 1977) are also left out of annotation since they go beyond our scope. Bridging holds between two elements in which the second element is interpreted by an inferential process ("bridge") from the first, but the two elements do not corefer. A bridging inference between *l'Escola Coral* 'the Choral School' and *els alumnes* 'the students' (11) is triggered by the definite article in the latter NP.

[14] Given the length of some discourse segments, in the examples of discourse deixis coreferent mentions are underlined in order to distinguish them clearly from their antecedent.

[15] We are replacing *cada una* 'each' with the coreferent candidate *EDF y Mitsubishi* 'EDF and Mitsubishi.' In the English translation, an inversion of verb-subject order is required.

	MUC	ACE	MATE	AnCora-CO
1. Elliptical pronouns			✔	✔
2. Clitic pronouns			✔	✔
3. Quoted speech	✔	✔	✔	✔
4. Possessives	✔	✔	✔	✔
5. Embedded NPs	✔	✔	✔	✔
6. Split antecedent			✔	✔
7. Referential versus attributive		✔	✔	✔
8. Generic versus specific		✔		✔
9. Metonymy		✔	✔	✔
10. Discourse deixis			✔	✔
11. Bound anaphora	✔		✔	
12. Bridging reference			✔	

Table 2.1: Coverage of different coreference coding schemes

(11) (Cat.) L'Orfeó Manresà posa en marxa el mes d'octubre *l'Escola Coral*. Es tracta d'un projecte destinat a despertar en *els alumnes* la passió pel cant coral.
'The Manresa Orfeo starts *the Choral School* in October. It is a project aimed at arousing among *the students* a passion for choral singing.'

2.4 Annotation scheme

Despite the existence of a few coreference annotation schemes, there is no standard as yet, a shortcoming largely accounted for by the complexities of the linguistic phenomenon (see Section 2.3). Due to space constraints, we will not go into detail about the various annotation schemes used in former annotation endeavours. Instead, Table 2.1 sums up three of the most widely-used existing schemes by showing whether or not they include (✔) the issues outlined in Section 2.3. The first two were used to encode the corpora for the MUC and ACE programs (HIRSCHMAN i CHINCHOR, 1997; DODDINGTON *et al.*, 2004); the MATE meta-scheme (DAVIES *et al.*, 1998; POESIO, 2004*b*) is different in that it is not linked with a specific corpus but constitutes a proposal for dialogue annotation with a wide range of potential tags from which the designer can build his own scheme. The final column in Table 2.1 sets the coding scheme used in the AnCora-CO corpora against the other two, highlighting the arguments put forward in the previous section.

The MUC and ACE schemes depend to a great extent on the evaluation tasks for which the corpora were originally developed, which makes them either inconsistent or limited from a linguistic point of view. In contrast, the flexibility offered

by the MATE meta-scheme and its proposals for languages other than English has prompted us to adopt it—taking into account subsequent revisions and implementations (POESIO, 2004*b*; POESIO i ARTSTEIN, 2008)—as the model on which we base our annotation scheme for the AnCora-CO corpora.[16] Our aim is for AnCora-CO to be used to train/test coreference resolution systems as well as for linguistic enquiries and research on coreference. Consequently, the annotated features in our scheme are not only thought of as useful learning features but also linguistically motivated.

In order to set limits to render the annotation task feasible, we elected to restrict it to:

(a) Coreference links, ruling out any consideration of bound anaphora and bridging relations.

(b) NP reference. Other expressions like clauses and sentences are only encoded if they are subsequently referred to by an NP.

The task of coreference annotation involves two types of activities: marking of mentions and marking of coreference chains (entities).

2.4.1 Mentions

Given that AnCora already contains other annotation layers, the starting point for the marking of mentions was the existing rich hierarchical syntactic annotation. On the one hand, identifying mention candidates by using the output of the manual syntactic annotation freed coders from worrying about the exact boundaries of NPs. On the other hand, the existing syntactic tags constrained some decisions concerning coreference annotation. Nine types of syntactic nodes were eligible to be mentions:

(a) sn (NP)

(b) grup.nom (nominal group in a conjoined NP)

(c) relatiu (relative pronoun)

(d) d (possessive determiner)[17]

(e) p (possessive pronoun)[17]

(f) v (verb)[18]

(g) grup.verb (verbal group)

(h) S (clause)

(i) sentence

[16] http://clic.ub.edu/corpus/webfm_send/15

[17] The POS of possessive determiners and pronouns contains the entity corresponding to the possessor, the entire NP contains the entity corresponding to the thing(s) possessed.

[18] Verb nodes can only be a mention if they contain an incorporated clitic. The intention in annotating the verb is actually annotating the reference of the clitic, and this applies in Spanish only.

Units (a)-(f) are those considered as potential mentions in a coreference chain, while units (g)-(i) are only included in a coreference chain if they are subsequently referred to by one of the other units. To indicate whether (a)-(f) mentions are referential or not, the attribute *entityref* is specified with one out of five possible values (the absence of the attribute is one of the values). The first three values identify the set of referential mentions, i.e., mention candidates to participate in a coreference link (see Section 2.4.2 below).

1. Named entity ("ne"). The concept of named entity (NE) has its origins in the Named Entity Recognition and Classification tasks, an offspring of Information Extraction systems, and it is still central today in the NLP field, being a core element in the ACE competition. Information about NEs in AnCora comes from existing semantic annotations (BORREGA *et al.*, 2007), where NEs are defined as those nouns whose referent is unique and unambiguous, e.g., *Obama*; *onze del matí* '11 am.' They fall into six semantic types: person, organization, location, date, number and others (publications, prizes, laws, etc.). Coreference annotation takes into account weak NEs, as these are the ones marked at the NP level.[19] They are either NPs containing a proper noun (e.g., *Los Angeles*; *la ciudad de Los Angeles* 'the city of Los Angeles'), or definite NPs whose head is a common noun modified by a national or a relational adjective (e.g., *el gobierno vasco* 'the Basque government').

2. Specific ("spec"). Specific mentions corefer with an NE and have the form of an anaphoric pronoun (12-a) or a full NP that contains no proper noun or trigger word (12-b).

   (12) a. (Sp.) Klebánov[entityref="ne"] manifestó que ∅[entityref="spec"] no puede garantizar el éxito al cien por cien.
   'Klebánov stated that *(he)* cannot guarantee 100% success.'
   b. (Cat.) En un sentit similar s'ha manifestat Jordi Pujol[entityref="ne"] ... *El president* [entityref="spec"] ha recordat ...
   'To a similar effect Jordi Pujol voiced his opinion ... *The president* recalled ...'

3. Non-named entity ("nne"). This value identifies mentions that refer to an entity with no specific name (13); that is, referential mentions which are neither "spec" nor "ne."

   (13) (Sp.) *La expansión de la piratería en el Sudeste de Asia* puede destruir las economías de la región.
   '*The extension of piracy in South-East Asia* could destroy the economies of the region.'

[19] Strong NEs correspond strictly to the POS level (nouns, e.g., *Los Angeles*).

4. Lexicalized ("lex"). Lexicalized mentions are non-referential mentions that are part of a set phrase or idiom (14-a), including clitics inherent in pronominal verbs (14-b).

   (14) a. (Sp.) Dar *las gracias*.
   'To give *thanks*.'
   b. (Cat.) Passar-*les* magres.
   'To have a hard time.'[20]

5. No *entityref* attribute indicates that the mention is non-referential (and other than lexicalized). It can be an attributive NP (15-a), a nominal predicate (15-b), an appositive phrase, a predicative complement (15-c), a negated NP (15-d), an interrogative pronoun (15-e), a measure NP (15-f), or the Catalan partitive pronoun *en*.

   (15) a. (Sp.) Sistema de *educación*.
   '*Education* system.'
   b. (Sp.) La hipótesis de la colisión era *la más probable*.
   'The collision hypothesis was *the most likely*.'
   c. (Sp.) Julio Valdés fue elegido como *el quinto mejor futbolista de Centroamérica*.
   'Julio Valdés was chosen as *the fifth best football player in Central America*.'
   d. (Sp.) No se les exige *ninguna prueba de capacitación*.
   '*No proficiency test* is required of them.'
   e. (Sp.) Las dudas sobre *quien* ganará las elecciones.
   'The doubts as to *who* is going to win the elections.'
   f. (Sp.) Andrés Palop estará *cuatro meses* de baja.
   'Andrés Palop will be on leave for *four months*.'

A second attribute, *homophoricDD*, is meant to identify Halliday and Hasan's (1976) homophoric definite descriptions, which are proper-noun-like and generic definite NPs that refer to something in the cultural context or world view, e.g., (Cat.) *la ira* 'the anger', *l'actualitat* 'the present time', *les dones* 'women.' A test for homophoricity is whether the mention can be the first mention of an entity in a text, i.e., requiring no previous introduction. The NEs that appear in newspaper articles are usually assumed to be already hearer-old and, if not, they are accompanied by a relative clause or an appositive. Therefore, this attribute is not specified for NEs, but only for mentions that are entityref="nne" and definite (introduced by the definite article). Notice that, unlike English, generic NPs in Spanish and Catalan are introduced by the definite article.

[20]The original version with the inherent clitic is untranslatable into English.

The third attribute specific to mentions is *title*. It is assigned the value "yes" if the mention is part of a newspaper headline or subheading.

2.4.2 Coreference chains

Coreferent mentions are assigned an *entity* attribute whose value specifies an entity number ("entity#"). Hence, the collection of mentions referring to the same discourse entity all have the same entity number. Our set of coreference relations restricts those proposed in MATE to three, which correspond to the three values that the *coreftype* attribute can take. A *coreftype* is specified for all mentions coreferent with a previous one. Additionally, mentions linked either by a discourse deixis or a predicative relation contain a *corefsubtype* attribute with semantic information. The different coreference types and subtypes are now commented and exemplified, thus highlighting the range of relations contemplated by our scheme. The annotation guidelines explicitly went for high precision at the expense of possibly low recall: coders were told to avoid any dubious link.

- Identity ("ident"). This tag marks referential mentions that point to the same discourse entity as a previous mention in the text. What we call a "transitivity test" is performed to check whether an identity relation holds between two mentions: if mention A can occupy the slot that mention B occupies in the text with no change in meaning, then A and B corefer.[21] Table 2.2 shows a sample of mention pairs from different entities (AnCora-CO-Es). The sixth row illustrates an instance of a split antecedent that results from the union of Entity 1 and Entity 4.

- Discourse deixis ("dx"). Following the terminology proposed by WEBBER (1988), this tag is used for mentions that corefer with a previous verb, clause, or one or more sentences (16). The set of possible antecedents is given by the underlying syntactic annotations: mentions of types (g)-(i), i.e., verbs, clauses, and sentences.

(16) a. (Sp.) *Un pirata informático consiguió robar los datos de 485.000 tarjetas de crédito* ... El robo fue descubierto.
'*A hacker managed to steal data from 485,000 credit cards.* ... The theft was uncovered.'

b. (Cat.) El 1966, *la monja va vomitar sang*. El fet es va repetir al cap de sis mesos.
'In 1966, *the nun brought up blood*. The incident recurred six months later.'

c. (Sp.) El jefe de las Fuerzas Armadas de Sudáfrica, el general Nyanda, afirmó en su primera visita oficial a Angola que *las*

[21] The transitivity test extends to all the mentions in the same entity so that if mention A corefers with mention B, and mention B corefers with mention C, then it is possible to replace mention C by mention A with no change in meaning, and vice versa.

Entity	Mention$_a$	Mention$_b$	Mention$_b$ form
Entity1	*el cuarto socio de CGO* 'the fourth partner of CGO'	*IJM Corporation Berhad*	Proper noun
Entity2	*Buenos Aires*	*la capital argentina* 'the Argentinian capital'	Definite NP
Entity3	*acciones* 'shares'	*acciones* 'shares'	Bare NP
Entity4	*tres de las empresas de CGO* 'three of the companies of CGO'	*ø*	Elliptical subject
Entity1+4	los socios de CGO 'the partners of CGO'	*que* 'that'	Relative pronoun
Entity5	*Ecuador*	*le* 'it'	Third person pronoun
Entity6	*mi equipo* 'my team'	*nosotros* 'we'	First person pronoun
Entity7	*Emil Zapotek*	*un superhombre capaz de ganar* 'a superman capable of winning'	Indefinite NP
Entity8	*Barça*	*ganarle* 'beat|them'	Clitic pronoun

Table 2.2: Sample of mentions with an identity link (AnCora-CO-Es)

> *Fuerzas Armadas de este país "consiguieron destruir buena parte de las fuerzas convencionales de UNITA*." El general sudafricano hizo estas declaraciones.
> 'The head of the Armed Forces of South Africa, general Nyanda, stated on his first official visit to Angola that *the Armed Forces of this country "managed to destroy a large part of UNITA's conventional forces*." The South African general made these declarations.'

Since discourse-deictic mentions can make reference to different aspects of a previous discourse segment, they take a *corefsubtype* attribute, which can be of three types:

- Token (16-a). The mention refers to the same event-token (i.e., same spatial and temporal coordinates) as the previous segment.
- Type (16-b). The mention refers to an event of the same type as the segment, but not the same token.
- Proposition (16-c). The mention refers to the segment as a linguistic object, i.e., the proposition itself.

Existing corpora annotated with discourse deixis are small (ECKERT i STRUBE, 2000; NAVARRETTA, 2007). The coreference annotation in the ongoing OntoNotes project—developing three large corpora for English, Chinese and Arabic—includes discourse deixis but only considers the heads of VPs as possible antecedents (PRADHAN *et al.*, 2007*b*). This is the most straightforward solution, but it might fail to capture the precise extension of the

antecedent. The coreference annotation of AnCora-CO is done on top of the already existing syntactic annotation, which conditions in some cases the coreference annotation because a discourse segment can be considered to be the antecedent from a linguistic perspective, but the segment might not be a syntactic constituent.

- Predicative ("pred"). This tag identifies attributes of a mention that are expressed by a nominal predicate (17-a), an appositive phrase (17-b,c), or a parenthetical phrase (17-d). These relations are not coreferential, but keeping track of predicative information can be helpful when training a computational coreference resolution system, since an attribute often adds information by renaming or further defining a mention. Besides, as stated previously, by including predicative links we give users the chance to decide whether or not they prefer to collapse the distinction between coreference and predication.

(17) a. (Sp.) Unión Fenosa Inversiones es *una empresa del grupo español Unión Fenosa*.
'Unión Fenosa Inversiones is *a company in the Spanish group Unión Fenosa*.'

b. (Cat.) Han demanat una entrevista amb el conseller d'Indústria, *Antoni Subirà*.
'They have asked for an interview with the Minister of Industry, *Antoni Subirà*.'

c. (Cat.) Hi podrà participar tothom, actuant com a moderadora Montserrat Clua, *membre de la facultat d'Antropologia de la Universitat Autònoma de Barcelona*.
'Everybody will be able to participate. Montserrat Clua, *a member of the faculty of Anthropology at the Autonoma University of Barcelona*, will act as a moderator.'

d. (Sp.) Los ministros de Defensa de la Unión Europea *(UE)* celebrarán el próximo lunes en Bruselas una conferencia.
'The Ministers of Defence of the European Union *(EU)* will be attending a conference in Brussels next Monday.'

Predicative link types contain a *corefsubtype* that indicates a semantic distinction, specifying whether the attribution is:

- Definite. A definite attribution occurs in both equative and identificational clauses, in which a defining feature of the subject is described (17-b,d). It might be expressed by a proper noun, a phrase introduced by the definite article, or a bare NP.[22]

[22] In Spanish and Catalan, unlike English, equative appositive and copular phrases often omit the definite article.

```
<sn arg="arg0" entity="entity1" entityref="ne" func="suj" ne="organization" tem="agt">
    <spec gen="f" num="s">
      <d gen="f" lem="el" num="s" postype="article" wd="La"/>
    </spec>
    <grup.nom gen="f" num="s">
      <n entityref="ne" gen="c" lem="Comisión_Europea" ne="organization" num="c"
     postype="proper" sense="16:cs1" wd="Comisión_Europea"/>
    </grup.nom>
  </sn>
  <grup.verb>
    <v els="a2" lem="anunciar" mood="indicative" num="s" person="3"
   postype="main" tense="past" wd="anunció"/>
  </grup.verb>
  <sadv arg="argM" func="cc" functype="temporal" tem="tmp">
    <grup.adv>
      <r lem="hoy" wd="hoy"/>
    </grup.adv>
  </sadv>
  <S arg="arg1" clausetype="completive" func="cd" impersonal="no" tem="pat">
    <conj conjunctiontype="subordinating">
      <c lem="que" postype="subordinating" wd="que"/>
    </conj>
    <sn arg="arg0" coreftype="ident" elliptic="yes" entity="entity1" entityref="spec"
   func="suj" tem="agt"/>
    <grup.verb>
      <v lem="haber" num="s" person="3" postype="auxiliary" tense="present" wd="ha"/>
      <v els="a2" lem="recibir" num="s" postype="main" wd="recibido"/>
    </grup.verb>
    <sn arg="arg1" entityref="nne" func="cd" homophoricDD="yes" tem="pat">
      <spec gen="f" num="s">
        <d gen="f" lem="el" num="s" postype="article" wd="la"/>
      </spec>
      <grup.nom gen="f" num="s">
        <n gen="f" lem="notificación" num="s" postype="common" sense="16:05388391"
       wd="notificación"/>
      </grup.nom>
    </sn>
  </S>
  <f lem="." punct="period" wd="."/>
```

Figure 2.1: The XML file format exemplified with the sentence *La Comisión Europea anunció hoy que ∅ ha recibido la notificación* 'The European Commission announced today that (it) received the notification.' Notice that the bold entity number "entity1" marks the identity coreference relation between *la Comisión Europea* 'the European Commission' and an elliptical subject ('it')

- Indefinite. A characterizing but non-identificative feature of the mention (17-a,c) is expressed.

Negated or modal predicates (18) are not annotated since they either say what the mention is not, or provide a description dependent on a subjective perspective.

(18) (Sp.) Andalucía no es *propiedad del PSOE*.
'Andalusia is not *the property of the PSOE*.'

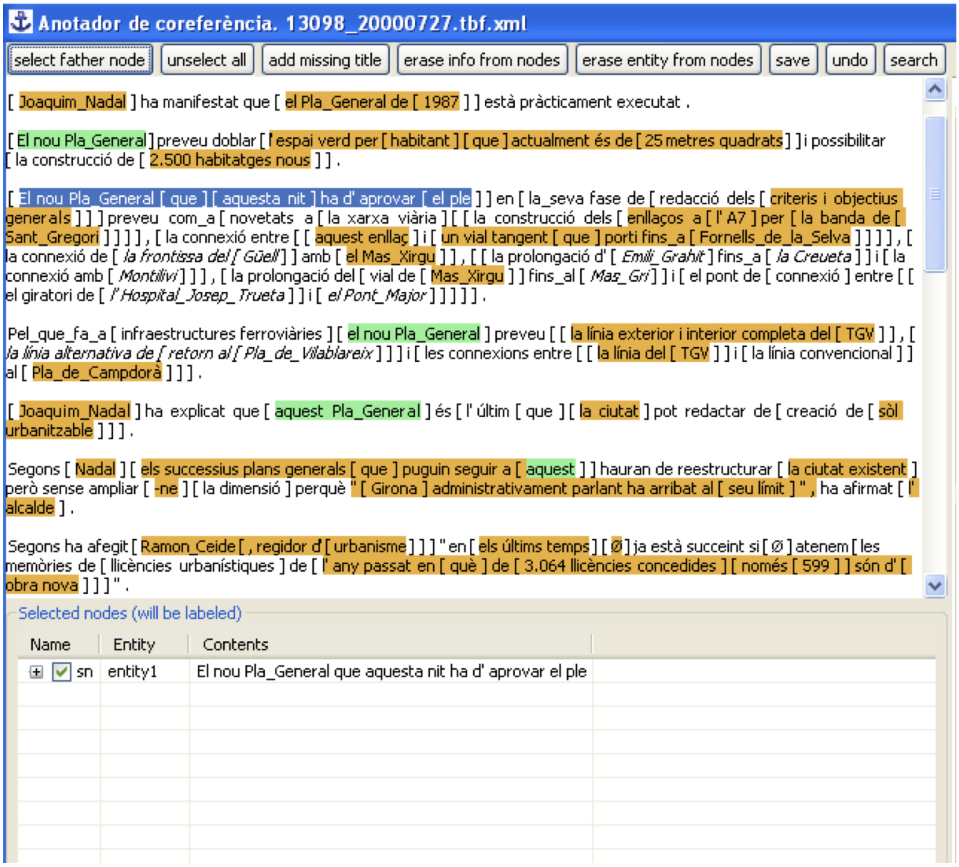

Figure 2.2: Left screenshot of the coreference annotation tool in AnCoraPipe

2.5 Annotation tool

The corpus was created using AnCoraPipe (BERTRAN *et al.*, 2008), an annotation tool developed at the University of Barcelona for the purpose of accommodating and unifying the attribute-value pairs of each coding level. To this end, the tool uses the same XML data storage format for each stage (Fig. 2.1). Given that the previous annotation layers of AnCora were already encoded in an in-line fashion, AnCoraPipe employs this format, unlike other similar tools, such as MMAX2 (MÜLLER i STRUBE, 2006), which support standoff markup. Although the advantages of standoff coding are well known (IDE, 2000), especially in resolving the conflict of overlapping hierarchies of data elements, the conversion of AnCora-CO to a standoff data architecture remains a project for the future.

The tool efficiently handles annotation on multiple linguistic levels, and coders can easily switch from one level to another (e.g., to correct mistakes found in another layer). In this way, the required annotation time is reduced and the integration of the coders' work is seamless. The corpora in the local machine are associated with a server so that, as soon as an annotator modifies a file, the latter is uploaded

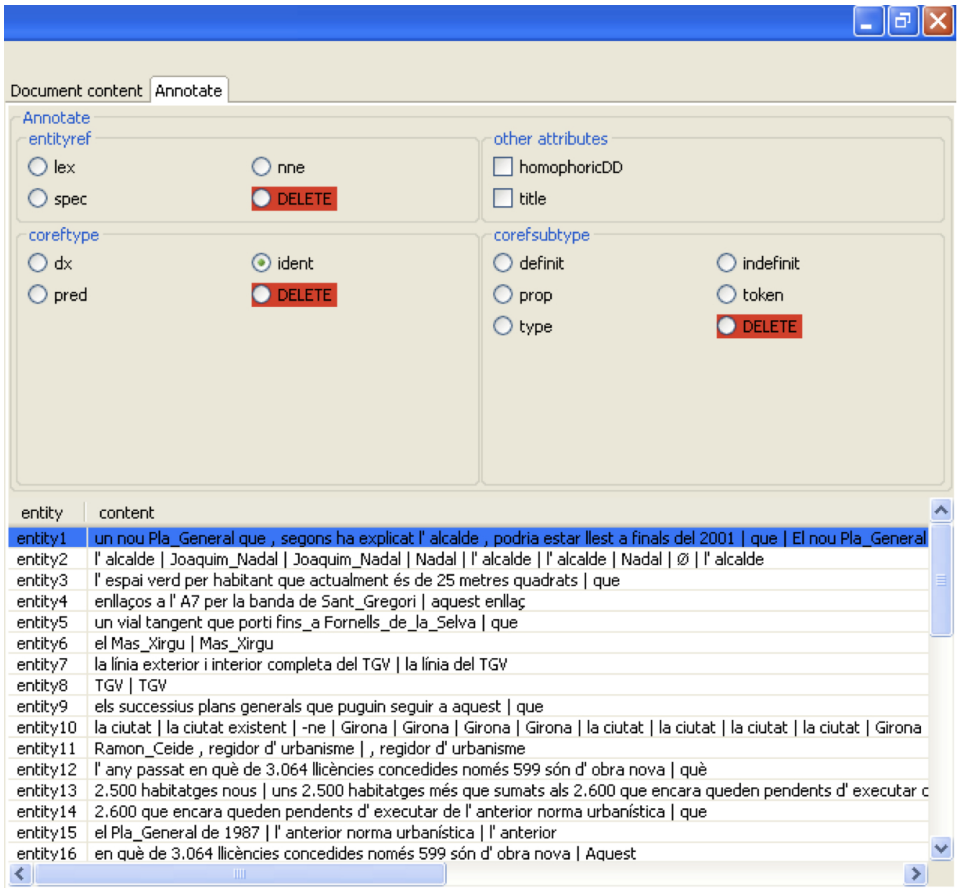

Figure 2.3: Right screenshot of the coreference annotation tool in AnCoraPipe

to the server before other users add further annotations.

AnCoraPipe provides an additional tool for coreference annotation that makes the process faster and more user-friendly (Figs. 2.2, 2.3). Mentions that are annotated with an entity number appear highlighted in the text in different colours. Attribute-values can easily be added, changed, or removed. Fig. 2.2 shows the left side of the screen and Fig. 2.3 shows the right side of the screen. The screen is divided into four panels:

Top left (Fig. 2.2, top). The raw text contained in one file (i.e., one newspaper article).

Bottom left (Fig. 2.2, bottom). The selected syntactic nodes being labelled.

Top right (Fig. 2.3, top). The attributes-values information.

Bottom right (Fig. 2.3, bottom). The collection of annotated multi-mention entities.

In Fig. 2.2, the NP *El nou Pla General que aquesta nit ha d'aprovar el ple* is the mention currently considered as a potential coreferent mention. In order to

add it to an entity (i.e., a coreference chain), the coder clicks on the corresponding entity in the window bottom right (Fig. 2.3). The values of the rest of attributes for this mention are selected in the window top right (Fig. 2.3). All mentions with the same entity number ("entity1" in this example) corefer.

A total of seven annotators contributed to the process of enriching AnCora with coreference information, although throughout the process the average number of coders working at any given time was never more than three. They were all graduates or final-year undergraduates of linguistics, and were paid for their work. The annotation process was divided into two stages: a first pass in which all mention attributes and coreference links were coded, and a second pass in which the newly annotated files were revised.

2.6 Distributional statistics

This section provides distributional statistics for the coreference tags in the corpora, which are very similar for the two languages under consideration. AnCora-CO-Es (422,887 tokens) contains 134,247 NPs, of which 24,380 (18.16%) are not marked as referential mentions. AnCora-CO-Ca (385,831 tokens) contains 122,629 NPs, of which 24,906 (20.31%) are non-referential. Table 2.3 shows the distribution of mentions, and provides details of the number of mentions sorted by POS. We distinguish between isolated, first, and subsequent mentions. It emerges that about 1/2 of the mentions are isolated, 1/6 are first mentions, and 1/3 are subsequent mentions. Coreferential mentions are split into pronouns (1/3) and full NPs (2/3).

The number of entities, including those containing a single mention, is 89,206 in Spanish, and 81,386 in Catalan. The distribution of coreftype and corefsubtype tags over mentions marked as coreferent is presented in Table 2.4. These are pairwise links, which means that 17,884 non-single-mention entities include 45,909 links (AnCora-CO-Es), and 16,545 non-single-mention entities include 41,959 links (AnCora-CO-Ca). Table 2.5 shows the distribution of entities according to their size (i.e., the number of mentions they contain).

These statistics reveal interesting linguistic issues which could open up many avenues for future research. Notice, for instance, the high percentage of definite NPs that are isolated or first mentions, which confirms the findings of the studies conducted by FRAURUD (1990) and POESIO i VIEIRA (1998) in Swedish and English, respectively. The number of first-mention definites in Spanish and Catalan is even higher (see RECASENS *et al.* (2009*a*) for a more detailed exploration).

2.7 Inter-annotator agreement

There is widespread agreement on the fact that coders' judgments in semantic and pragmatic annotation tasks such as coreference are very subjective and, consequently, that the resulting annotations need to be tested for reliability. To this end, inter-annotator agreement is assessed. Consistency can only be achieved if the

POS	AnCora-CO-Es			AnCora-CO-Ca		
	Isolated[a]	First[b]	Subsequent[c]	Isolated[a]	First[b]	Subsequent[c]
Pronoun						
Personal	0.26	0.08	1.76	0.35	0.07	2.20
Elliptical	0.44	0.18	5.74	0.44	0.13	4.98
Relative	1.41	0.01	4.43	0.68	0.01	4.97
Demonstrative	0.13	0.06	0.15	0.19	0.06	0.16
Subtotal	**2.25**	**0.33**	**12.08**	**1.67**	**0.28**	**12.31**
Full NP						
Bare common N	10.42	0.71	0.90	11.68	0.87	0.99
Bare proper N	5.72	2.02	5.76	5.71	1.79	4.93
Indefinite	5.06	1.43	0.88	5.01	1.60	0.97
Definite	17.73	7.19	11.46	19.32	7.63	12.78
Demonstrative	0.59	0.16	0.96	0.69	0.15	1.17
Possessive[d]	2.14	0.41	0.58	–	–	–
Numeral	2.96	0.37	0.22	2.58	0.39	0.15
Subtotal	**44.62**	**12.28**	**20.76**	**44.99**	**12.44**	**20.99**
Coordinated	2.77	0.35	0.29	3.28	0.38	0.31
Misc.	3.49	0.35	0.41	2.93	0.40	0.03
Total	**53.13**	**13.32**	**33.55**	**52.88**	**13.49**	**33.63**

[a] Isolated mentions are entities with a single mention in the text.
[b] First mentions are the first reference to a multi-mention entity.
[c] Subsequent mentions are references to a multi-mention entity other than first mentions.
[d] Possessive NPs are always preceded by the definite article in Catalan, so they are included in the count of definites.

Table 2.3: Distribution of mentions according to POS and chain position (%)

Coreftype	Corefsubtype	AnCora-CO-Es	AnCora-CO-Ca
Identity		89.11	91.42
Discourse deixis		2.50	2.35
	Token	1.88	1.74
	Type	0.22	0.34
	Proposition	0.40	0.27
Predicative		8.39	6.23
	Definite	6.48	4.90
	Indefinite	1.91	1.33

Table 2.4: Distribution of coreftype and corefsubtype tags (%)

Entity size	AnCora-CO-Es	AnCora-CO-Ca
1 mention	79.95	79.66
2 mentions	11.15	11.25
3-5 mentions	6.46	6.64
6-10 mentions	1.72	1.77
$>$ 10 mentions	0.72	0.68

Table 2.5: Distribution of entity tags according to number of mentions (%)

coding instructions are appropriate for the data, and annotators understand how to apply them. A reliability study on a sample of the corpus makes it possible to pinpoint both the strengths and weaknesses of the coding scheme, and make the necessary changes before proceeding to the annotation of the entire corpus.

Different agreement coefficients have been used by the discourse processing community, but there is no standardized metric for agreement on coreference. In their survey, ARTSTEIN i POESIO (2008) point out the main problems in using percent agreement and the kappa coefficient (SIEGEL i CASTELLAN, 1988; CARLETTA, 1996). On the one hand, percent agreement does not yield values that can be compared across studies, since some agreement is due to chance, and the amount of chance agreement is affected by two factors that vary from one study to another:

(a) The number of categories (the fewer categories, the higher the agreement expected by chance).

(b) The distribution of items among categories (the more common a category, the higher the agreement expected by chance).

On the other hand, kappa is corrected for chance agreement, but it is not appropriate for all types of agreement because it assumes that all disagreements are equal. A third coefficient, alpha (α), overcomes the two previous limitations by being both chance-corrected and weighted (KRIPPENDORFF, 2004 [1980]).

2.7.1 Reliability study

In this section we present a reliability study on the annotation scheme presented in Section 2.4, as applied to data from AnCora-CO. Given the high cost of conducting such studies, time, budget and personnel constraints prompted us to limit the scope of the experiment to the core tag of the coreference coding scheme (the coreftype attribute) and to data from the Spanish corpus as a representative sample. Taking into account that most work on reference is limited to pronominal anaphors and has used kappa, we were mainly interested in analyzing to what extent coders agreed on assigning identity versus non-coreference relations for both pronominal and non-pronominal NPs. Specifically, we set out to:

1. Examine the coverage and tag definitions of the coding scheme.

2. Test the adequacy and clarity of the annotation guidelines.
3. Identify cases raising significant issues, with a view to establishing a typology of sources of disagreement.

The results show that the annotation of AnCora-CO is reliable to an acceptable degree.[23] Thus, the corpora can serve as a valuable language resource on which to base studies of coreference in Catalan and Spanish, as well as reference on a more general level.

2.7.1.1 Subjects

Six volunteer undergraduates (with no previous experience in corpus annotation) and two linguistics graduates (two of the annotators who had worked on the corpus) participated in the experiment, all of them students at the University of Barcelona and native bilingual Spanish-Catalan speakers.

2.7.1.2 Materials

A total of four newspaper texts from the AnCora-CO-Es corpus were used: two[24] (838 tokens, 261 mentions) in the training stage, and the other two[25] (1,147 tokens, 340 mentions) in the testing stage. In both cases, the second text was more complex than the first one, being longer and including a higher number of ambiguities and discourse-deictic relations. Given the shortage of time, the chosen texts were short, but each one included at least two instances of every link type.

2.7.1.3 Tools

The annotations were performed on three computers with Windows XP using the PALinkA annotation tool (ORASAN, 2003).[26]

2.7.1.4 Procedure

The experiment was run in four ninety-minute sessions: two training sessions and two testing sessions. Annotators were given the set of mentions (NPs) and had to decide for each of them whether it was coreferent or not. If so, the appropriate value for the coreftype attribute had to be selected, in addition to the entity. During the first two sessions, coders familiarized themselves with the annotation tool and guidelines, and feedback was provided to each of them after the mock annotation

[23] It is common practice among researchers in Computational Linguistics to consider 0.8 the absolute minimum value of α to accept for any serious purpose (ARTSTEIN i POESIO, 2008).

[24] Files 11177_20000817 and 16468_20000521.

[25] Files 17704_20000522 (Text 1, 62 coreferent mentions) and 17124_0001122 (Text 2, 88 coreferent mentions).

[26] At the time of the experiment, AnCoraPipe (the annotation tool that was used for the actual annotation) was not ready yet.

Mention	Coder A	Coder B	Coder C	Coder D	Coder E	Coder F	Coder G	Coder H
m0	1	1	1	1	1	1	1	1
m1	1	1	1	1	1	1	1	1
m2	3	3	3	3	3	3	3	3
m3	1	1	1	1	1	1	1	1
m4	1	1	1	1	1	1	1	1
m5	1	1	1	1	1	1	1	1
m6	1	1	1	1	1	1	1	1
m7	1	1	1	1	1	1	1	1
m8	1	1	1	1	1	1	1	1
m9	1	1	1	1	1	1	1	1
m10	1	1	1	4	1	4	1	1

Table 2.6: Partial agreement matrix for Text 1 (Each value identifies a different link type: 1 = non-coreference; 2 = discourse deixis; 3 = predicative; 4 = identity)

of two texts. In the last two sessions, they annotated the two test texts separately from each other.

2.7.1.5 Results

ARTSTEIN i POESIO (2008) make the point that coreference encoding differs from other annotation tasks in that coders do not assign a specific label to each category but create collections of coreferent mentions. PASSONNEAU (2004) proposes using the emerging coreference chains (i.e., entities) as the labels, and recommends the MASI (Measuring Agreement on Set-valued Items) distance metric (PASSONNEAU, 2006) to allow for partial agreement. In our experiment, it turned out that disagreements emerged from different decisions on the link type assigned to a mention rather than on the same mention being assigned to different entities by different coders. As a result, we decided to use two agreement values to separate the two aspects: (a) link type (treating non-coreference as a type), and (b) entity number. The first was measured by Krippendorff's α, as disagreements are not all alike. The second was measured by kappa, as there was no need for weighted agreement.

To measure link type, the four coreftype links (non-coreference, identity, predicative, discourse deixis) were used as the possible labels that could be assigned to each mention. PASSONNEAU (2004) employs a coder-by-item agreement matrix where the row labels are the items (mentions), the column labels are the coders, and the cell contents indicate the value that a specific coder assigned to a specific item. This kind of matrix was used to enter the results of the experiment (Table 2.6), where a numerical value identifies each link type. Krippendorff's α was computed with the freely available KALPHA macro written for SPSS (HAYES i KRIPPENDORFF, 2007), yielding the following results: $\alpha = .85$ ([.828,.864] 95% CI) for Text 1, and $\alpha = .89$ ([.872,.896] 95% CI) for Text 2. Krippendorff's α ranges between -1 and 1, where 1 signifies perfect agreement and 0 signifies no

	Non-coref	Dx	Pred	Ident
Non-coref	690.71	4.14	7.29	34.86
Dx	4.14	9.71	.00	.14
Pred	7.29	.00	89.14	.57
Ident	34.86	.14	.57	331.43

Table 2.7: Observed coincidence matrix (Text 1)

	Non-coref	Dx	Pred	Ident
Non-coref	446.81	8.50	58.89	222.80
Dx	8.50	.15	1.12	4.23
Pred	58.89	1.12	7.67	29.32
Ident	222.80	4.23	29.32	110.64

Table 2.8: Expected coincidence matrix (Text 1)

difference from chance agreement (rather than no agreement).

To measure entity number, a coder-by-item agreement matrix similar to the previous one (Table 2.6) was used, but in this case the row labels only contain the mentions that were linked by an identity or predicative relation,[27] and the cells contain the entity number they were assigned. In fact, there was just a single case in which coders disagreed (see (19), below, in Section 2.7.2). Thus, high kappa values were obtained: κ=.98 for Text 1, and κ=1 for Text 2.

2.7.1.6 Discussion

In the observed coincidence matrix (Table 2.7) for link type, the disagreements between observers cluster around the diagonal containing perfect matches. The expected coincidence matrix (Table 2.8) can be interpreted as what would be expected under conditions of chance. The delta matrix (Table 2.9) shows how α weights the coincidences: a mismatch between non-coreference and discourse deixis is less penalized—subtler decision—than one between non-coreference and predicative, while the stiffest penalization is for disagreement between non-coreference and identity, which are the labels at either end of the spectrum.

Even now, according to ARTSTEIN i POESIO (2008), it is "the lack of consensus on how to interpret the values of agreement coefficients" that accounts for "the reluctance of many in Computational Linguistics to embark on reliability studies." In his work, KRIPPENDORFF (2004 [1980]) suggests α=.8 as a threshold value, which is supported by more recent efforts (ARTSTEIN i POESIO, 2005). In both texts, we obtained an α coefficient above .8, which is high enough to claim good reliability as far as the four-way distinction between

[27] Discourse-deictic relations were left out from the quantitative study since coders only received the set of NPs as possible mentions. They had free choice to select the discourse segment antecedents. For the qualitative analysis on this respect, see Section 2.7.2 below.

	Non-coref	Dx	Pred	Ident
Non-coref	.00	141000.25	185761.00	439569.00
Dx	141000.25	.00	3080.25	82656.25
Pred	185761.00	3080.25	.00	53824.00
Ident	439569.00	82656.25	53824.00	.00

Table 2.9: Delta matrix (Text 1)

non-coreference : identity : discourse deixis : predicative

is concerned. Contrary to our expectations, Text 2 yields a higher reliability score, which is possibly due to the different size: Text 1 contains 152 mentions, and Text 2 contains 188 mentions. Even though the second text contains some tricky coreference relations, it also contains many clear cases of non-coreferential mentions, which increase the intercoder agreement. The high alpha results from the fact that the coding guidelines define precisely the relations covered by each link type, thus separating identity from predicative links and ruling out less well-defined relations such as bridging. Likewise, the preference expressed in the annotation manual for excluding any link in case of doubt or ambiguity—as in cases of only partial identity—accounts for the almost full agreement obtained for entity number. The guidelines discuss how to deal with recurrent non-prototypical cases of coreference, although there will always be new cases not covered by the manual, or obscure to coders, which account for the margin up to full agreement.

The general pattern is that two out of the eight coders (which can already be seen from the agreement matrix, Table 2.6) account for the majority of disagreements, and they do not deviate in the same direction, which provides further support of the validity of the guidelines as most mistakes can be attributed to certain coders' poorer understanding of the annotation task. If these two outliers are removed and α is recomputed with the other six coders, the results improve up to $\alpha = .87$ ([.857,.898] 95% CI) for Text 1, and $\alpha = .90$ ([.882,.913] 95% CI) for Text 2. The remaining disagreements are broken down in the next section.

2.7.2 Sources of disagreement

A reliability study informs about intercoder agreement and also enables disagreements to be analyzed so as to improve data reliability and better understand the linguistic reality. Detecting sources of unreliability provides an insight into weaknesses of the annotation guidelines, the complexity of the linguistic phenomenon under analysis and the aptitude of the coders. After computing the exact reliability agreement, we compared qualitatively the output of the eight coders, going into more detail than with the four-way distinction of the coreftype attribute. We grouped the major sources of disagreement under seven headings.

1. Different metonymic interpretation. Metonymy accounts for the only case of disagreement on entity number, giving rise to two different plausible in-

terpretations. The qualitative analysis uncovered the fact that *las dos delegaciones* '*the two delegations*' in (19) can be linked either with the two spokesmen involved (*the head of the Armed Forces of South Africa* and *general Joao de Matos*) or with the two respective countries (*South Africa* and *Angola*).

(19) (Sp.) El jefe de las Fuerzas Armadas de Sudáfrica, el general Nyanda, afirmó en su primera visita oficial a Angola que ... En su visita, el general Nyanda estuvo acompañado por el general Joao de Matos ... Según fuentes próximas al Ministerio de Defensa, durante las conversaciones entre *las dos delegaciones* ...
'The head of the Armed Forces of South Africa, general Nyanda, stated during his first official visit to Angola that ... On his visit, general Nyanda was accompanied by general Joao de Matos ... According to sources close to the Ministry of Defence, during the conversations between *the two delegations* ...'

2. Violations of the maximal NP principle. Three disagreements were caused by the coders' failure to notice that the reference of an embedded mention (20-b) coincided with the entire NP mention (20-a), thus disagreeing on the mention annotated as coreferent. (20-a) and (20-b) show the two different mentions selected as coreferent with *su reinado* 'his reign' by different coders. It is only the entire NP (20-a) that should be annotated as coreferent since it refers to Juan Carlos I's reign by its duration, thus coinciding with the element referenced by *reinado* 'reign.'

(20) a. (Sp.) *los veinticinco años de reinado de Juan Carlos I*
'*the twenty-five years of reign of Juan Carlos I*'
b. (Sp.) los veinticinco años de *reinado de Juan Carlos I*
'the twenty-five years of *reign of Juan Carlos I*'

3. Idiolinks. Each coder produced at least one link that none of the rest did. They were usually the result of unclear coreference or a bridging relation. In (21) the reference of the two mentions overlaps but is not identical: what the King has promoted is just a part of what the King has done for the country. Even if coders were told not to annotate cases of bridging, it seems it was hard for them to ignore these relations if they saw one.

(21) (Sp.) *lo que el Rey ha impulsado ... lo que el Rey ha hecho por el país*
'*what the King has promoted ... what the King has done for the country*'

4. Referential versus attributive NPs. The divide between referential and at-

tributive mentions turned out to be unclear to two coders, who linked the two attributive NPs in (22).

(22) (Sp.) misión de *paz* ... fuerzas de *paz*
'*peacekeeping* mission ... *peacekeeping* forces'

5. Discourse deixis. Even though the computation of Krippendorff's α only took into account whether annotators agreed on the mentions in a discourse-deictic relation (and they did in the four cases found in the test texts), the qualitative analysis revealed that they did not always coincide in the syntactic node of the discourse segment chosen as antecedent. In the following example, half of the coders selected the previous clause (23-a) while the other half selected the entire previous sentence (23-b) as the antecedent of the mention *estas declaraciones* 'these declarations.'

(23) a. (Sp.) El jefe de las Fuerzas Armadas de Sudáfrica, el general Nyanda, afirmó en su primera visita oficial a Angola que *las Fuerzas Armadas de este país "consiguieron destruir buena parte de las fuerzas convencionales de UNITA*". El general sudafricano hizo estas declaraciones.
'The head of the Armed Forces of South Africa, general Nyanda, stated on his first official visit to Angola that *the Armed Forces of this country "managed to destroy a large part of UNITA's conventional forces*". The South African general made these declarations.'

b. (Sp.) *El jefe de las Fuerzas Armadas de Sudáfrica, el general Nyanda, afirmó en su primera visita oficial a Angola que las Fuerzas Armadas de este país "consiguieron destruir buena parte de las fuerzas convencionales de UNITA*". El general sudafricano hizo estas declaraciones.
'*The head of the Armed Forces of South Africa, general Nyanda, stated on his first official visit to Angola that the Armed Forces of this country "managed to destroy a large part of UNITA's conventional forces*". The South African general made these declarations.'

6. Missed links. Each coder missed one or two links. The reason for this was either sheer oversight or because s/he did not recognize them as an instance of coreference.

7. Misunderstandings. The two coders that produced the most naïve annotations were misled by cases where two NP heads matched semantically (i.e., same string) but not referentially.

(24) (Sp.) El próximo *envío* de tropas sudafricanas en el marco de la Misión de la ONU en el vecino Congo ... el *envío* de 5.500 cascos azules para la RDC
'The next *dispatch* of South-African troops within the framework of the UN Mission in the neighbouring Congo' ... 'the *dispatch* of 5,500 blue berets for the DRC'

In a nutshell, most of the problems can be attributed to a lack of training (i.e., familiarity with the guidelines) on the part of the coders, as well as oversights or ambiguities left unresolved in the discourse itself. After carrying out the study, it became clear that the guidelines were clear and adequate, and that, assuming coders go through a period of training, many disagreements that were just a matter of error or misapplication could be resolved through revision. Therefore, we decided that a two-pass procedure was required to annotate the whole corpus: each text was annotated twice by two different coders, thus always revising the links from the first pass and checking for missing ones. The qualitative analysis of the sources of disagreements shows the subtleties of the task of coreference annotation and hence the need for qualified linguists to build a reliable language resource, in line with KILGARRIFF (1999).

2.8 Conclusions

We presented the enrichment of the AnCora corpora with coreference information, which heralded the advent of the AnCora-CO corpora. The Spanish and Catalan corpora constitute a language resource that can be used for both studying coreference relations and training automatic coreference resolution systems. The AnCora-CO corpora contain coreference annotations for Spanish and Catalan conjoined with morphological, syntactic and semantic information, thus making it possible to rely on a wide range of learning features to train computational systems. This can be especially helpful for coreference resolution, which is known to be a very challenging task, given that many sources of knowledge come into play. In this respect, AnCora-CO opens new avenues for carrying out research on the way coreference links—both between pronouns and full NPs—are established by language users.

Given the subjectivity of discourse phenomena like coreference, there is a need to understand the linguistic problem so as to produce thorough and useful annotation guidelines (ZAENEN, 2006). This was our main guiding principle. The annotation scheme designed to annotate coreference draws on the MATE/GNOME/ARRAU scheme, but restricting it to coreference. Special attention was paid to finding a balance between the hypothetical requirements of a machine-learning coreference resolution system and the way in which the linguistic reality allows itself to be encoded. The key to our approach lies in three central factors. First, relations are split into three kinds: identity of reference, discourse deixis, and predication. Other re-

lations such as bridging are not included in order to keep a consistent definition of coreference. Second, what is meant by "identity of reference" is clarified with the help of real examples to reduce ambiguities to a great extent. The transitivity test is used as an indicator of coreference. Third, mentions are individually tagged with three attributes containing information (entity reference, homophoric definite description, title) that can be used to group mentions into referential/non-referential, and first/subsequent mentions.

The quality of the scheme was assessed by computing intercoder agreement in a reliability study with eight coders. We used kappa to measure agreement on entity number, and Krippendorff's alpha to test the reliability of the link type attribute, which is the core of the scheme as it separates non-coreferential from identity, predicative and discourse-deictic mentions. Once a mention was chosen as being coreferent, the choice of entity was widely agreed upon. The high inter-annotator agreement demonstrated the reliability of the annotation, whereas the dissection of the disagreements served to suggest a typology of errors and determine the best procedure to follow. We leave for future work a large-scale reliability study that explores further issues such as the identification of antecedents in discourse deixis.

In order to do the markup, the AnCoraPipe annotation tool was customised to meet our needs. Since the XML format enables the corpora to be easily extended with new annotation levels, AnCora-CO can be further extended to include, for example, coding of nominal argumental structures, discourse markers, etc. In addition, we intend to convert the current in-line annotation to a standoff format. By developing the AnCora-CO corpora we have provided Spanish and Catalan with two new language resources.

Acknowledgements This work was supported by the FPU Grant (AP2006-00994) from the Spanish Ministry of Education and Science, and the Lang2World (TIN2006-15265-C06-06) and Ancora-Nom (FFI2008-02691-E/FILO) projects. Special thanks to Mariona Taulé for her invaluable advice, Manuel Bertran for customising the AnCoraPipe annotation tool, and the annotators who participated in the development of AnCora-CO and the reliability study: Oriol Borrega, Isabel Briz, Irene Carbó, Sandra García, Iago González, Esther López, Jesús Martínez, Laura Muñoz, Montse Nofre, Lourdes Puiggròs, Lente Van Leeuwen, and Rita Zaragoza. We are indebted to three anonymous reviewers for their comments on earlier versions of this work.

Part II

RESOLUCIÓ I AVALUACIÓ DE LA COREFERÈNCIA

CAPÍTOL 3

A Deeper Look into Features for Coreference Resolution

Marta Recasens* and Eduard Hovy**

*University of Barcelona
**USC Information Sciences Institute

Published in S. Lalitha Devi, A. Branco, and R. Mitkov (eds.),
Anaphora Processing and Applications (DAARC 2009), LNAI 5847:29-42,
Springer-Verlag, Berlin Heidelberg

Abstract All automated coreference resolution systems consider a number of features, such as head noun, NP type, gender, or number. Although the particular features used is one of the key factors for determining performance, they have not received much attention, especially for languages other than English. This paper delves into a considerable number of pairwise comparison features for coreference, including old and novel features, with a special focus on the Spanish language. We consider the contribution of each of the features as well as the interaction between them. In addition, given the problem of class imbalance in coreference resolution, we analyze the effect of sample selection. From the experiments with TiMBL (Tilburg Memory-Based Learner) on the AnCora corpus, interesting conclusions are drawn from both linguistic and computational perspectives.

Keywords Coreference resolution · Machine learning · Features

3.1 Introduction

Coreference resolution, the task of identifying which mentions in a text point to the same discourse entity, has been shown to be beneficial in many NLP applications

such as Information Extraction (MCCARTHY i LEHNERT, 1995), Text Summarization (STEINBERGER *et al.*, 2007), Question Answering (MORTON, 1999), and Machine Translation. These systems need to identify the different pieces of information concerning the same referent, produce coherent and fluent summaries, disambiguate the references to an entity, and solve anaphoric pronouns.

Given that many different types of information—ranging from morphology to pragmatics—play a role in coreference resolution, machine learning approaches (SOON *et al.*, 2001; NG i CARDIE, 2002*b*) seem to be a promising way to combine and weigh the relevant factors, overcoming the limitations of constraint-based approaches (LAPPIN i LEASS, 1994; MITKOV, 1998), which might fail to capture global patterns of coreference relations as they occur in real data. Learning-based approaches decompose the task of coreference resolution into two steps: (i) classification, in which a classifier is trained on a corpus to learn the probability that a pair of NPs are coreferent or not; and (ii) clustering, in which the pairwise links identified at the first stage are merged to form distinct coreference chains.

This paper focuses on the classification stage and, in particular, on (i) the features that are used to build the feature vector that represents a pair of mentions,[1] and (ii) the selection of positive and negative training instances. The choice of the information encoded in the feature vectors is of utmost importance as they are the basis on which the machine learning algorithm learns the pairwise coreference model. Likewise, given the highly skewed distribution of coreferent vs. non-coreferent classes, we will consider whether sample selection is helpful. The more accurate the classification is, the more accurate the clustering will be.

The goal of this paper is to provide an in-depth study of the pairwise comparison stage in order to decrease as much as possible the number of errors that are passed on to the second stage of coreference resolution. Although there have been some studies in this respect (URYUPINA, 2007; BENGTSON i ROTH, 2008; HOSTE, 2005), they are few, oriented to the English or Dutch language, and dependent on poorly annotated corpora. To our knowledge, no previous studies compared systematically a large number of features relying on gold standard corpora, and experiments with sample selection have been only based on small corpora. For the first time, we consider the degree of variance of the learnt model on new data sets by reporting confidence intervals for precision, recall, and F-score measures.

The paper is organized as follows. In the next section, we review previous work. In Section 3.3, we list our set of 47 features and argue the linguistic motivations behind them. These features are tested by carrying out different machine learning experiments with TiMBL in Section 3.4, where the effect of sample selection is also assessed. Finally, main conclusions are drawn in Section 3.5.

[1] This paper restricts to computing features over a pair of mentions—without considering a more global approach—hence *pairwise comparison features*.

3.2 Previous work

Be it in the form of hand-crafted heuristics or feature vectors, what kind of knowledge is represented is a key factor for the success of coreference resolution. Although theoretical studies point out numerous linguistic factors relevant for the task, computational systems usually rely on a small number of shallow features, especially after the burst of statistical approaches. In learning-based approaches, the relative importance of the factors is not manually coded but inferred automatically from an annotated corpus. Training instances for machine learning systems are feature vectors representing two mentions (m_1 and m_2) and a label ("coreferent" or "non-coreferent") allowing the classifier to learn to predict, given a new pair of NPs, whether they do or do not corefer.

The feature set representing m_1 and m_2 that was employed in the decision tree learning algorithm of SOON *et al.* (2001) has been taken as a starting point by most subsequent systems. It consists of only 12 surface-level features (all boolean except for the first): (i) sentence distance, (ii) m_1 is a pronoun, (iii) m_2 is a pronoun, (iv) string match (after discarding determiners), (v) m_2 is a definite NP, (vi) m_2 is a demonstrative NP, (vii) number agreement, (viii) WordNet semantic class agreement,[2] (ix) gender agreement, (x) both m_1 and m_2 are proper nouns (capitalized), (xi) m_1 is an alias of m_2 or vice versa, and (xii) m_1 is an apposition to m_2. The strongest indicators of coreference turned out to be string match, alias, and appositive.

NG i CARDIE (2002*b*) expanded the feature set of SOON *et al.* (2001) from 12 to a deeper set of 53, including a broader range of lexical, grammatical, and semantic features such as substring match, comparison of the prenominal modifiers of both mentions, animacy match, WordNet distance, whether one or both mentions are pronouns, definite, embedded, part of a quoted string, subject function, and so on. The incorporation of additional knowledge succeeds at improving performance but only after manual feature selection, which points out the importance of removing irrelevant features that might be misleading. Surprisingly, however, some of the features in the hand-selected feature set do not seem very relevant from a linguistic point of view, like string match for pronominal mentions.

More recent attempts have explored some additional features to further enrich the set of NG i CARDIE (2002*b*): backward features describing the antecedent of the candidate antecedent (YANG *et al.*, 2004), semantic information from Wikipedia, WordNet and semantic roles (PONZETTO i STRUBE, 2006), and most notably, URYUPINA's (2007) thesis, which investigates the possibility of incorporating sophisticated linguistic knowledge into a data-driven coreference resolution system trained on the MUC-7 corpus. Her extension of the feature set up to a total of 351 nominal features (1096 boolean/continuous) leads to a consistent improvement in the system's performance, thus supporting the hypothesis that complex

[2]Possible semantic classes for an NP are *female, male, person, organization, location, date, time, money, percent*, and *object*.

linguistic factors of NPs are a valuable source of information. At the same time, however, URYUPINA (2007) recognizes that by focusing on the addition of sophisticated features she overlooked the resolution strategy and some phenomena might be over-represented in her feature set.

BENGTSON i ROTH (2008) show that with a high-quality set of features, a simple pairwise model can outperform systems built with complex models on the ACE dataset. This clearly supports our stress on paying close attention to designing a strong, linguistically motivated set of features, which requires a detailed analysis of each feature individually as well as of the interaction between them. Some of the features we include, like modifiers match, are also tested by BENGTSON i ROTH (2008) and, interestingly, our ablation study comes to the same conclusion: almost all the features help, although some more than others.

HOSTE's (2005) work is concerned with optimization issues such as feature and sample selection, and she stresses their effect on classifier performance. The study we present is in line with URYUPINA (2007), BENGTSON i ROTH (2008) and HOSTE (2005), but introduces a number of novelties. First, the object language is Spanish, which presents some differences as far as coreference is concerned. Second, we use a different corpus, AnCora, which is twenty times as large as MUC and, unlike ACE, it includes a non-restricted set of entity types. Third, the coreference annotation of the AnCora corpus sticks to a linguistic definition of the identity relationship more accurate than that behind the MUC or ACE guidelines. Fourth, we do not rely on the (far from perfect) output of preprocessing modules but take advantage of the gold standard annotations in the AnCora corpus in order to focus on their real effect on coreference resolution.

3.3 Pairwise comparison features

The success of machine learning systems depends largely on the feature set employed. Learning algorithms need to be provided with an adequate representation of the data, that is to say, a representation that includes the "relevant" information, to infer the best model from an annotated corpus. Identifying the constraints on when two NPs can corefer is a complex linguistic problem that remains still open. Hence, there is a necessity for an in-depth study of features for coreference resolution from both a computational and a linguistic perspective. This section makes a contribution in this respect by considering a total of 47 features, making explicit the rationale behind them.

- **Classical features** (Table 3.1). The features that have been shown to obtain better results in previous works (SOON *et al.*, 2001; NG i CARDIE, 2002*b*; LUO *et al.*, 2004) capture the most basic information on which coreference depends, but form a reduced feature set that does not account for all kinds of coreference relations.
  - PRON_m_1 and PRON_m_2 specify whether the mentions are pronouns

Feature	Definition	Value
PRON_m_1	m_1 is a pronoun	true, false
PRON_m_2	m_2 is a pronoun	true, false
HEAD_MATCH	Head match	true, false, ?[a]
WORDNET_MATCH	EuroWordNet match	true, false, ?[a]
NP_m_1	m_1 NP type	common, proper, article, indefinite, possessive, relative, demonstrative, numeral, interrogative, personal, exclamative
NP_m_2	m_2 NP type	common, proper, article, indefinite, possessive, relative, demonstrative, numeral, interrogative, personal, exclamative
NE_m_1	m_1 NE type	person, organization, location, date, number, other, null
NE_m_2	m_2 NE type	person, organization, location, date, number, other, null
NE_MATCH	NE match	true, false, ?[b]
SUPERTYPE_MATCH	Supertype match	true, false, ?[a]
GENDER_AGR	Gender agreement	true, false
NUMBER_AGR	Number agreement	true, false
ACRONYM	m_2 is an acronym of m_1	true, false, ?[c]
QUOTES	m_2 is in quotes	true, false
FUNCTION_m_1	m_1 function	subject, d-obj, i-obj, adjunct, prep-obj, attribute, pred-comp, agent, sent-adjunct, no function
FUNCTION_m_2	m_2 function	subject, d-obj, i-obj, adjunct, prep-obj, attribute, pred-comp, agent, sent-adjunct, no function
COUNT_m_1	m_1 count	#times m_1 appears in the text
COUNT_m_2	m_2 count	#times m_2 appears in the text
SENT_DIST	Sentence distance	#sentences between m_1 and m_2
MENTION_DIST	Mention distance	#NPs between m_1 and m_2
WORD_DIST	Word distance	#words between m_1 and m_2

[a] Not applicable. This feature is only applicable if neither m_1 nor m_2 are pronominal or conjoined.
[b] Not applicable. This feature is only applicable if both mentions are NEs.
[c] Not applicable. This feature is only applicable if m_2 is an acronym.

Table 3.1: Classical features

Feature	Definition	Value
ELLIP_m_1	m_1 is an elliptical pronoun	true, false
ELLIP_m_2	m_2 is an elliptical pronoun	true, false
GENDER_PRON	Gender agreement restricted to pronouns	true, false, ?
GENDER_MASCFEM	Gender agreement restricted to masc./fem.	true, false, ?
GENDER_PERSON	Gender agreement restricted to persons	true, false, ?
ATTRIBa_m_1	m_1 is attributive type A	true, false
ATTRIBa_m_2	m_2 is attributive type A	true, false
ATTRIBb_m_1	m_1 is attributive type B	true, false
ATTRIBb_m_2	m_2 is attributive type B	true, false

Table 3.2: Language-specific features

Feature	Definition	Value
NOMPRED_m_1	m_1 is a nominal predicate	true, false
NOMPRED_m_2	m_2 is a nominal predicate	true, false
APPOS_m_1	m_1 is an apposition	true, false
APPOS_m_2	m_2 is an apposition	true, false
PRONTYPE_m_1	m_1 pronoun type	elliptical, 3-person, non-3-person, demonstrative, possessive, indefinite, numeric, other, ?
PRONTYPE_m_2	m_2 pronoun type	elliptical, 3-person, non-3-person, demonstrative, possessive, indefinite, numeric, other, ?
EMBEDDED	m_2 is embedded in m_1	true, false
MODIF_m_1	m_1 has modifiers	true, false
MODIF_m_2	m_2 has modifiers	true, false

Table 3.3: Corpus-specific features

Feature	Definition	Value
FUNCTION_TRANS	Function transition	100 different values (e.g., subject_subject, subject_d-obj)
COUNTER_MATCH	Counter match	true, false, ?
MODIF_MATCH	Modifiers match	true, false, ?
VERB_MATCH	Verb match	true, false, ?
NUMBER_PRON	Number agreement restricted to pronouns	true, false, ?
TREE-DEPTH_m_1	m_1 parse tree depth	#nodes in the parse tree from m_1 up to the top
TREE-DEPTH_m_2	m_2 parse tree depth	#nodes in the parse tree from m_2 up to the top
DOC_LENGTH	Document length	#tokens in the document

Table 3.4: Novel features

since these show different patterns of coreference, e.g., gender agreement is of utmost importance for pronouns but might be violated by non-pronouns (HOSTE, 2005).

- HEAD_MATCH is the top classical feature for coreference, since lexical repetition is a common coreference device.
- WORDNET_MATCH uses the Spanish EuroWordNet[3] and is true if any of the synset's synonyms of one mention matches any of the synset's synonyms of the other mention.
- NP type plays an important role because not all NP types have the same capability to introduce an entity into the text for the first time, and not all NP types have the same capability to refer to a previous mention in the text.
- The fact that in newspaper texts there is usually at least one person and a location about which something is said accounts for the relevance of the NE type feature, since NE types like *person* and *organization* are more likely to corefer and be coreferred than others.
- SUPERTYPE_MATCH compares the first hypernym of each mention found in EuroWordNet.
- As a consequence of the key role played by gender and number in anaphora resolution, GENDER_AGR and NUMBER_AGR have been inherited by coreference systems. See below, however, for finer distinctions.
- The rationale behind QUOTES is that a mention in quotes identifies a mention that is part of direct speech, e.g., if it is a first- or second- person pronoun, its antecedent will be found in the immediate discourse.

• **Language-specific features** (Table 3.2). There are some language-specific issues that have a direct effect on the way coreference relations occur in a language. In the case of Spanish, we need to take into account elliptical subjects, grammatical gender, and nouns used attributively.

- There is a need to identify elliptical pronouns in Spanish because, unlike overt pronouns, they get their number from the verb, have no gender, and always appear in subject position, as shown in (1), where the elliptical subject pronoun is marked with ø and with the corresponding pronoun in brackets in the English translation.

  (1) Klebánov manifestó que ø no puede garantizar el éxito al cien por cien.
  'Klebánov stated that *(he)* cannot guarantee 100% success.'

[3]Nominal synsets are part of the semantic annotation of AnCora. EuroWordNet covers 55% of the nouns in the corpus.

  - Since Spanish has grammatical gender, two non-pronominal nouns with different gender might still corefer, e.g., *el incremento* 'the increase' (masc.) and *la subida* 'the rise' (fem.). Gender agreement is an appropriate constraint only for pronouns.
  - GENDER_MASCFEM does not consider those NPs that are not marked for gender (e.g., elliptical pronouns, companies).
  - GENDER_PERSON separates natural from grammatical gender by only comparing the gender if one of the mentions is an NE-person.[4]
  - Attributive NPs[5] are non-referential, hence non-markables. ATTRIBa and ATTRIBb identify two Spanish constructions where these NPs usually occur:

    Type A. Common, singular NPs following the preposition *de* 'of', e.g., *educación* 'education' in *sistema de educación* 'education system.'

    Type B. Proper nouns immediately following a generic name, e.g., *Mayor* 'Main' in *calle Mayor* 'Main Street'.

- **Corpus-specific features** (Table 3.3). The definition of coreference in the AnCora corpus differs from that of the MUC and ACE corpora in that it separates identity from other kinds of relation such as apposition, predication, or bound anaphora. This is in line with VAN DEEMTER i KIBBLE's (2000) criticism of MUC. Predicative and attributive NPs do not have a referential function but an attributive one, qualifying an already introduced entity. They should not be allowed to corefer with other NPs. Consequently, the use we make of nominal-predicate and appositive features is the opposite to that made by systems trained on the MUC or ACE corpora (SOON *et al.*, 2001; LUO *et al.*, 2004). Besides, the fact that AnCora contains gold standard annotation from the morphological to the semantic levels makes it possible to include additional features that rely on such rich information.

  - We employ NOMPRED to filter out predicative mentions.
  - We employ APPOS to filter out attributively used mentions.
  - Gold standard syntactic annotation makes it possible to assess the efficacy of the EMBEDDED and MODIF features in isolation from any other source of error. First, a nested NP cannot corefer with the embedding one. Second, depending on the position a mention occupies in the coreference chain, it is more or less likely that it is modified.

- **Novel features** (Table 3.4). We suggest some novel features that we believe relevant and that the rich annotation of AnCora enables.

[4] Animals are not included since they are not explicitly identitifed as NEs.

[5] *Attributively* used NPs qualify another noun.

– FUNCTION_TRANS is included because although FUNCTION_m_1 and FUNCTION_m_2 already encode the function of each mention separately, there may be information in their joint behaviour.[6] E.g., *subject_subject* can be relevant since two consecutive subjects are likely to corefer:

(2) [...] explicó *Alonso, quien anunció la voluntad de Telefónica Media de unirse a grandes productoras iberoamericanas*. Por otra parte, *Alonso* justificó el aplazamiento.
'[...] explained *Alonso, who announced the will of Telefónica Media to join large Latin American production companies*. On the other hand, *Alonso* justified the postponement.'

– COUNTER_MATCH prevents two mentions that contain a different numeral to corefer (e.g., *134 millones de euros* '134 million euros' and *194 millones de euros* '194 million euros'), as they point to a different number of referents.

– Modifiers introduce extra information that might imply a change in the referential scope of a mention (e.g., *las elecciones generales* 'the general elections' and *las elecciones autonómicas* 'the regional elections'). Thus, when both mentions are modified, the synonyms and immediate hypernym of the head of each modifying phrase are extracted from EuroWordNet for each mention. MODIF_MATCH is true if one of them matches between the two mentions.

– The verb, as the head of the sentence, imposes restrictions on its arguments. In (3), the verb *participate* selects for a volitional agent, and the fact that the two subjects complement the same verb hints at their coreference link. VERB_MATCH is true if either the two verbal lemmas or any synonym or immediate hypernym from EuroWordNet match.

(3) *Un centenar de artistas* participará en el acto [...] el acto se abrirá con un brindis en el que participarán *todos los protagonistas de la velada*.
'*One hundred artists* will participate in the ceremony [...] the ceremony will open with a toast in which *all the protagonists of the evening gathering* will participate.'

– NUMBER_PRON is included since non-pronominal mentions that disagree in number might still corefer.

– DOC_LENGTH can be helpful since the longer the document, the more coreferent mentions, and a wider range of patterns might be allowed.

[6] The idea of including conjoined features is also exploited by BENGTSON i ROTH (2008) and LUO *et al.* (2004).

	Training set	Test set
# Words	298,974	23,022
# Entities	64,421	4,893
# Mentions	88,875	6,759
# NEs	25,758	2,023
# Nominals	53,158	4,006
# Pronominals	9,959	730

Table 3.5: Characteristics of the AnCora-Es datasets

3.4 Experimental evaluation

This section describes our experiments with the features presented in Section 3.3 as well as with different compositions of the training and test data sets. We finally assess the reliability of the most appropriate pairwise comparison model.

Data. The experiments are based on the AnCora-Es corpus (RECASENS i MARTÍ, 2010), a corpus of newspaper and newswire articles. It is the largest Spanish corpus annotated, among other levels of linguistic information, with PoS tags, syntactic constituents and functions, named entities, nominal WordNet synsets, and coreference links.[7] We split randomly the freely available labelled data into a training set of 300k words and a test set of 23k words. See Table 3.5 for a description.

Learning algorithm. We use TiMBL, the Tilburg memory-based learning classifier (DAELEMANS i BOSCH, 2005), which is a descendant of the *k*-nearest neighbor approach. It is based on analogical reasoning: the behavior of new instances is predicted by extrapolating from the similarity between (old) stored representations and the new instances. This makes TiMBL particularly appropriate for training a coreference resolution model, as the feature space tends to be very sparse and it is very hard to find universal rules that work all the time. In addition, TiMBL outputs the information gain of each feature—very useful for studies on feature selection—and allows the user easily to experiment with different feature sets by obscuring specified features. Given that the training stage is done without abstraction but by simply storing training instances in memory, it is considerably faster than other machine learning algorithms.

We select parameters to optimize TiMBL on a held-out development set. The distance metric parameter is set to overlap, and the number of nearest neighbors (*k* parameter) is set to 5 in Section 3.4.1, and to 1 in Section 3.4.2.[8]

[7] AnCora is freely available from http://clic.ub.edu/corpus/en/ancora.

[8] When training the model on the full feature vectors, the best results are obtained when TiMBL uses 5 nearest neighbors for extrapolation. However, because of the strong skew in the class space, in some of the hill-climbing experiments we can only use 1 nearest neighbor. Otherwise, with 5 neighbors the majority of neighbors are of the negative class for all the test cases, and the positive class is never predicted (recall=0).

	Training set		Test set	
	Representative	Balanced	Representative	Balanced
Positive instances	105,920		8,234	
Negative instances	425,942	123,335	32,369	9,399

Table 3.6: Distribution of representative and balanced data sets

	Training set	Test set	P	R	F
Model A	Representative	Representative	84.73	73.44	78.68
Model B	Representative	Balanced	88.43	73.44	80.24
Model C	Balanced	Representative	66.28	80.24	72.60
Model D	Balanced	Balanced	83.46	87.32	85.34

Table 3.7: Effect of sample selection on performance

3.4.1 Sample selection

When creating the training instances, we run into the problem of class imbalance: there are many more negative examples than positive ones. Positive training instances are created by pairing each coreferent NP with all preceding mentions in the same coreference chain. If we generate negative examples for all the preceding non-coreferent mentions, which would conform to the real distribution, then the number of positive instances is only about 7% (HOSTE, 2005). In order to reduce the vast number of negative instances, previous approaches usually take only those mentions between two coreferent mentions, or they limit the number of previous sentences from which negative mentions are taken. Negative instances have so far been created only for those mentions that are coreferent. In a real task, however, the system must decide on the coreferentiality of all mentions.

In order to investigate the impact of keeping the highly skewed class distribution in the training set, we create two versions for each data set: a representative one, which approximates the natural class distribution, and a balanced one, which results from down-sampling negative examples. The total number of negatives is limited by taking only 5 non-coreferent mentions randomly selected among the previous mentions (back to the beginning of the document). The difference is that in the balanced sample, non-coreferent mentions are selected for each coreferent mention, whereas in the representative sample they are selected for all mentions in the document. See Table 3.6 for statistics of the training and test sets.

Combining each training data set with each test set gives four possible combinations (Table 3.7) and we compute the performance of each of the models. The output of the experiments is evaluated in terms of precision (P), recall (R) and F-score (F). Although the best performance is obtained when testing the model on the balanced sample (models B and D), making a balanced test set involves knowledge about the different classes in the test set, which is not available in non-experimental situations. Therefore, being realistic, we must carry out the evaluation on a data

set that follows the natural class distribution. We focus our attention on models A and C.

Down-sampling on the training set increases R but at the cost of a too dramatic decrease in P. Because of the smaller number of negative instances in the training, it is more likely for an instance to be classified as positive, which harms P and F. As observed by HOSTE (2005), we can conclude that down-sampling does not lead to an increase in TiMBL, and so we opt for using model A.

3.4.2 Feature selection

This section considers the informativeness of the features presented in Section 3.3. We carry out two different feature selection experiments: (i) an ablation study, and (ii) a hill-climbing forward selection.

In the first experiment, we test each feature by running TiMBL on different subsets of the 47 features, each time removing a different one. The majority of features have low informativeness, as no single feature brings about a statistically significant loss in performance when omitted.[9] Even the removal of HEAD_MATCH, which is reported in the literature as one of the key features in coreference resolution, causes a statistically non-significant decrease of .15 in F. We conclude that some other features together learn what HEAD_MATCH learns on its own. Features that individually make no contribution are ones that filter referentiality, of the kind *ATTRIBb*_m_2, and ones characterising m_1, such as PRON_m_1. Finally, some features, in particular the distance and numeric measures, seem even to harm performance. However, there is a complex interaction between the different features. If we train a model that omits all features that seem irrelevant and harmful at the individual level, then performance on the test set decreases. This is in line with the ablation study performed by BENGTSON i ROTH (2008), who concludes that all features help, although some more than others.

Forward selection is a greedy approach that consists of incrementally adding new features—one at a time—and eliminating a feature whenever it causes a drop in performance. Features are chosen for inclusion according to their information gain values, as produced by TiMBL, most informative earliest. Table 3.8 shows the results of the selection process. In the first row, the model is trained on a single (the most informative) feature. From there on, one additional feature is added in each row; initial "-" marks the harmful features that are discarded (provide a statistically significant decrease in either P or R, and F). P and R scores that represent statistically significant gains and drops with respect to the previous feature vector are marked with an asterisk (*) and a dagger (†), respectively. Although F-score keeps rising steadily in general terms, informative features with a statistically significant improvement in P are usually accompanied by a significant decrease in R, and vice versa.

The results show several interesting tendencies. Although HEAD_MATCH is

[9] Statistical significance is tested with a one-way ANOVA followed by a Tukey's post-hoc test.

Feature vector	P	R	F	Feature vector	P	R	F
HEAD_MATCH	92.94	17.43	29.35	COUNTER_ MATCH	81.76	63.64	71.57
PRON_ m_2	57.58†	61.14*	59.30	MODIF_ m_1	81.08	64.67	71.95
ELLIP_ m_2	65.22*	53.04†	58.50	PRONTYPE_ m_1	81.70	64.84	72.30
-ELLIP_m_1	89.74*	34.09†	49.41	GENDER_ AGR	81.60	65.12	72.44
WORDNET_ MATCH	65.22	53.04	58.50	NOMPRED_ m_1	81.89	65.04	72.50
NE_ MATCH	65.22	53.04	58.50	GENDER_ PERSON	87.95*	64.78	74.61
-PRON_m_1	86.73*	38.74†	53.56	FUNCTION_ m_2	87.06	65.96	75.06
NUMBER_ PRON	69.04*	58.20*	63.16	FUNCTION_ m_1	85.88†	69.82*	77.02
-GENDER_PRON	86.64*	37.39†	52.24	QUOTES	85.83	70.11	77.18
VERB_ MATCH	80.31*	55.53†	65.66	COUNT_ m_2	85.62	70.73	77.47
SUPERTYPE_ MATCH	80.22	55.56	65.65	COUNT_ m_1	84.57	71.35	77.40
MODIF_ m_2	78.18	61.68*	68.96	NE_ m_1	83.82	72.48	77.74
NUMBER_ AGR	79.94	61.81	69.71	ACRONYM	83.99	72.46	77.80
ATTRIBb_ m_2	80.08	61.85	69.80	NE_ m_2	83.48	73.14	77.97
ATTRIBa_ m_2	80.14	61.84	69.81	NP_ m_2	82.81	73.55	77.91
ATTRIBa_ m_1	80.22	61.83	69.84	NP_ m_1	82.27	74.05	77.94
ATTRIBb_ m_1	80.23	61.82	69.83	FUNCTION_ TRANS	82.29	73.94	77.89
EMBEDDED	80.33	61.78	69.84	TREE-DEPTH_ m_2	80.54	72.98	76.57
GENDER_ MASCFEM	81.33	62.96	70.98	-TREE-DEPTH_ m_1	78.25†	72.52	75.27
APPOS_ m_1	81.46	62.96	71.02	-SENT_ DIST	78.17†	72.16	75.05
APPOS_ m_2	81.44	62.95	71.01	-DOC_ LENGTH	79.36*	70.36†	74.79
MODIF_ MATCH	81.35	63.10	71.08	MENTION_ DIST	79.52	72.10	75.63
NOMPRED_ m_2	81.38	63.37	71.26	WORD_ DIST	79.14	71.73	75.25
PRONTYPE_ m_2	81.70	63.59	71.52				

Table 3.8: Results of the forward selection procedure

the most relevant feature, it obtains a very low R, as it cannot handle coreference relationships involving pronouns or relations between full NPs that do not share the same head. Therefore, when PRON_ m_2 is added, R is highly boosted. With only these two features, P, R and F reach scores near the 60s. The rest of the features make a small—yet important in sum—contribution. Most of the features have a beneficial effect on performance, which provides evidence for the value of building a feature vector that includes linguistically motivated features. This includes some of the novel features we argue for, such as NUMBER_PRON and VERB_MATCH. Surprisingly, distance features seem to be harmful. However, if we train again the full model with the *k* parameter set to 5 and we leave out the numeric features, F does not increase but goes down. Again, the complex interaction between the features is manifested.

3.4.3 Model reliability

In closing this section, we would like to stress an issue to which attention is hardly ever paid: the need for computing the reliability of a model's performance. Because of the intrinsic variability in any data set, the performance of a model trained on one training set and tested on another will never be maximal. In addition to the two experiments varying feature and sample selection reported above, we actually carried out numerous other analyses of different combinations. Every change in the sample selection resulted in a change of the feature ranking produced by TiMBL.

For example, starting the hill-climbing experiment with a different feature would also lead to a different result, with a different set of features deemed harmful. Similarly, changing the test set will result in different performance of even the same model. For this reason, we believe that merely reporting system performances is not enough. It should become common practice to inspect evaluations taken over different test sets and to report the model's *averaged* performance, i.e., its F, R, and P scores, each bounded by confidence intervals.

To this end, we split randomly the test set into six subsets and evaluated each output. Then we computed the mean, variance, standard deviation, and confidence intervals of the six results of each P, R, and F-score. The exact performance of our pairwise comparison model for coreference (model A in Table 3.7) is 81.91±4.25 P, 69.57±8.13 R, and 75.12±6.47 F.

3.5 Conclusion

This paper focused on the classification stage of an automated coreference resolution system for Spanish. In the pairwise classification stage, the probability that a pair of NPs are or are not coreferent was learnt from a corpus. The more accurate this stage is, the more accurate the subsequent clustering stage will be. Our detailed study of the informativeness of a considerable number of pairwise comparison features and the effect of sample selection added to the few literature (URYUPINA, 2007; BENGTSON i ROTH, 2008; HOSTE, 2005) on these two issues.

We provided a list of 47 features for coreference pairwise comparison and discussed the linguistic motivations behind each one: well-studied features included in most coreference resolution systems, language-specific ones, corpus-specific ones, as well as extra features that we considered interesting to test. Different machine learning experiments were carried out using the TiMBL memory-based learner. The features were shown to be weakly informative on their own, but to support complex and unpredictable interactions. In contrast with previous work, many of the features relied on gold standard annotations, pointing out the need for automatic tools for ellipticals detection and deep parsing.

Concerning the selection of the training instances, down-sampling was discarded as it did not improve performance in TiMBL. Instead, better results were obtained when the training data followed the same distribution as the real-world data, achieving 81.91±4.25 P, 69.57±8.13 R, and 75.12±6.47 F-score. Finally, we pointed out the importance of reporting confidence intervals in order to show the degree of variance that the learnt model carries.

Acknowledgments We are indebted to M. Antònia Martí for her helpful comments.

This research was supported by the FPU Grant (AP2006-00994) from the Spanish Ministry of Education and Science, and the Lang2World (TIN2006-15265-C06-06) and Ancora-Nom (FFI2008-02691-E/FILO) projects.

CAPÍTOL 4

Coreference Resolution across Corpora: Languages, Coding Schemes, and Preprocessing Information

Marta Recasens* and Eduard Hovy**

*University of Barcelona
**USC Information Sciences Institute

Published in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1423–1432, Uppsala, Sweden

Abstract This paper explores the effect that different corpus configurations have on the performance of a coreference resolution system, as measured by MUC, B^3, and CEAF. By varying separately three parameters (language, annotation scheme, and preprocessing information) and applying the same coreference resolution system, the strong bonds between system and corpus are demonstrated. The experiments reveal problems in coreference resolution evaluation relating to task definition, coding schemes, and features. They also expose systematic biases in the coreference evaluation metrics. We show that system comparison is only possible when corpus parameters are in exact agreement.

4.1 Introduction

The task of coreference resolution, which aims to automatically identify the expressions in a text that refer to the same discourse entity, has been an increasing research topic in NLP ever since MUC-6 made available the first coreferentially annotated corpus in 1995. Most research has centered around the rules by which

mentions are allowed to corefer, the features characterizing mention pairs, the algorithms for building coreference chains, and coreference evaluation methods. The surprisingly important role played by different aspects of the corpus, however, is an issue to which little attention has been paid. We demonstrate the extent to which a system will be evaluated as performing differently depending on parameters such as the corpus language, the way coreference relations are defined in the corresponding coding scheme, and the nature and source of preprocessing information.

This paper unpacks these issues by running the same system—a prototype entity-based architecture called CISTELL—on different corpus configurations, varying three parameters. First, we show how much language-specific issues affect performance when trained and tested on English and Spanish. Second, we demonstrate the extent to which the specific annotation scheme (used on the same corpus) makes evaluated performance vary. Third, we compare the performance using gold-standard preprocessing information with that using automatic preprocessing tools.

Throughout, we apply the three principal coreference evaluation measures in use today: MUC, B^3, and CEAF. We highlight the systematic preferences of each measure to reward different configurations. This raises the difficult question of why one should use one or another evaluation measure, and how one should interpret their differences in reporting changes of performance score due to 'secondary' factors like preprocessing information.

To this end, we employ three corpora: ACE (DODDINGTON *et al.*, 2004), OntoNotes (PRADHAN *et al.*, 2007*a*), and AnCora (RECASENS i MARTÍ, 2010). In order to isolate the three parameters as far as possible, we benefit from a 100k-word portion (from the TDT collection) that is common to both ACE and OntoNotes. We apply the same coreference resolution system in all cases. The results show that a system's score is not informative by itself, as different corpora or corpus parameters lead to different scores. Our goal is not to achieve the best performance to date, but rather to expose various issues raised by the choices of corpus preparation and evaluation measure and to shed light on the definition, methods, evaluation, and complexities of the coreference resolution task.

The paper is organized as follows. Section 4.2 sets our work in context and provides the motivations for undertaking this study. Section 4.3 presents the architecture of CISTELL, the system used in the experimental evaluation. In Sections 4.4, 4.5, and 4.6, we describe the experiments on three different datasets and discuss the results. We conclude in Section 4.7.

4.2 Background

The bulk of research on automatic coreference resolution to date has been done for English and used two different types of corpus: MUC (HIRSCHMAN i CHINCHOR, 1997) and ACE (DODDINGTON *et al.*, 2004). A variety of learning-based systems have been trained and tested on the former (SOON *et al.*, 2001; URYUPINA,

2006), on the latter (CULOTTA *et al.*, 2007; BENGTSON i ROTH, 2008; DENIS i BALDRIDGE, 2009), or on both (FINKEL i MANNING, 2008; HAGHIGHI i KLEIN, 2009). Testing on both is needed given that the two annotation schemes differ in some aspects. For example, only ACE includes singletons (mentions that do not corefer) and ACE is restricted to seven semantic types.[1] Also, despite a critical discussion in the MUC task definition (VAN DEEMTER i KIBBLE, 2000), the ACE scheme continues to treat nominal predicates and appositive phrases as coreferential.

A third coreferentially annotated corpus—the largest for English—is OntoNotes (PRADHAN *et al.*, 2007*a*; HOVY *et al.*, 2006). Unlike ACE, it is not application-oriented, so coreference relations between all types of NPs are annotated. The identity relation is kept apart from the attributive relation, and it also contains gold-standard morphological, syntactic and semantic information.

Since the MUC and ACE corpora are annotated with only coreference information,[2] existing systems first preprocess the data using automatic tools (POS taggers, parsers, etc.) to obtain the information needed for coreference resolution. However, given that the output from automatic tools is far from perfect, it is hard to determine the level of performance of a coreference module acting on gold-standard preprocessing information. OntoNotes makes it possible to separate the coreference resolution problem from other tasks.

Our study adds to the previously reported evidence by STOYANOV *et al.* (2009) that differences in corpora and in the task definitions need to be taken into account when comparing coreference resolution systems. We provide new insights as the current analysis differs in four ways. First, STOYANOV *et al.* (2009) report on differences between MUC and ACE, while we contrast ACE and OntoNotes. Given that ACE and OntoNotes include some of the same texts but annotated according to their respective guidelines, we can better isolate the effect of differences as well as add the additional dimension of gold preprocessing. Second, we evaluate not only with the MUC and B^3 scoring metrics, but also with CEAF. Third, all our experiments use true mentions[3] to avoid effects due to spurious system mentions. Finally, including different baselines and variations of the resolution model allows us to reveal biases of the metrics.

Coreference resolution systems have been tested on languages other than English only within the ACE program (LUO i ZITOUNI, 2005), probably due to the fact that coreferentially annotated corpora for other languages are scarce. Thus there has been no discussion of the extent to which systems are portable across languages. This paper studies the case of English and Spanish.[4]

Several coreference systems have been developed in the past (CULOTTA *et al.*, 2007; FINKEL i MANNING, 2008; POON i DOMINGOS, 2008; HAGHIGHI i KLEIN,

[1] The ACE-2004/05 semantic types are person, organization, geo-political entity, location, facility, vehicle, weapon.

[2] ACE also specifies entity types and relations.

[3] The adjective *true* contrasts with *system* and refers to the gold standard.

[4] Multilinguality is one of the focuses of SemEval-2010 Task 1 (RECASENS *et al.*, 2010*b*).

2009; NG, 2009). It is not our aim to compete with them. Rather, we conduct three experiments under a specific setup for comparison purposes. To this end, we use a different, neutral, system, and a dataset that is small and different from official ACE test sets despite the fact that it prevents our results from being compared directly with other systems.

4.3 Experimental setup

4.3.1 System description

The system architecture used in our experiments, CISTELL, is based on the incrementality of discourse. As a discourse evolves, it constructs a model that is updated with the new information gradually provided. A key element in this model are the entities the discourse is about, as they form the discourse backbone, especially those that are mentioned multiple times. Most entities, however, are only mentioned once. Consider the growth of the entity *Mount Popocatépetl* in (1).[5]

(1) We have an update tonight on [this, the volcano in Mexico, they call El Popo]$_{m3}$ ... As the sun rises over [Mt. Popo]$_{m7}$ tonight, the only hint of the fire storm inside, whiffs of smoke, but just a few hours earlier, [the volcano]$_{m11}$ exploding spewing rock and red-hot lava. [The fourth largest mountain in North America, nearly 18,000 feet high]$_{m15}$, erupting this week with [its]$_{m20}$ most violent outburst in 1,200 years.

Mentions can be pronouns ($m20$), they can be a (shortened) string repetition using either the name ($m7$) or the type ($m11$), or they can add new information about the entity: $m15$ provides the supertype and informs the reader about the height of the volcano and its ranking position.

In CISTELL,[6] discourse entities are conceived as 'baskets': they are empty at the beginning of the discourse, but keep growing as new attributes (e.g., name, type, location) are predicated about them. Baskets are filled with this information, which can appear within a mention or elsewhere in the sentence. The ever-growing amount of information in a basket allows richer comparisons to new mentions encountered in the text.

CISTELL follows the learning-based coreference architecture in which the task is split into classification and clustering (SOON *et al.*, 2001; BENGTSON i ROTH, 2008) but combines them simultaneously. Clustering is identified with basket-growing, the core process, and a pairwise classifier is called every time CISTELL considers whether a basket must be clustered into a (growing) basket, which might contain one or more mentions. We use a memory-based learning classifier trained

[5] Following the ACE terminology, we use the term *mention* for an instance of reference to an object, and *entity* for a collection of mentions referring to the same object. Entities containing one single mention are referred to as *singletons*.

[6] 'Cistell' is the Catalan word for 'basket.'

with TiMBL (DAELEMANS i BOSCH, 2005). Basket-growing is done in four different ways, explained next.

4.3.2 Baselines and models

In each experiment, we compute three baselines (1, 2, 3), and run CISTELL under four different models (4, 5, 6, 7).

1. ALL SINGLETONS. No coreference link is ever created. We include this baseline given the high number of singletons in the datasets, since some evaluation measures are affected by large numbers of singletons.
2. HEAD MATCH. All non-pronominal NPs that have the same head are clustered into the same entity.
3. HEAD MATCH + PRON. Like HEAD MATCH, plus allowing personal and possessive pronouns to link to the closest noun with which they agree in gender and number.
4. STRONG MATCH. Each mention (e.g., m_{11}) is paired with previous mentions starting from the beginning of the document (m_1–m_{11}, m_2–m_{11}, etc.).[7] When a pair (e.g., m_3–m_{11}) is classified as coreferent, additional pairwise checks are performed with all the mentions contained in the (growing) entity basket (e.g., m_7–m_{11}). Only if *all* the pairs are classified as coreferent is the mention under consideration attached to the existing growing entity. Otherwise, the search continues.[8]
5. SUPER STRONG MATCH. Similar to STRONG MATCH but with a threshold. Coreference pairwise classifications are only accepted when TiMBL distance is smaller than 0.09.[9]
6. BEST MATCH. Similar to STRONG MATCH but following NG i CARDIE (2002*b*)'s best link approach. Thus, the mention under analysis is linked to the *most confident* mention among the previous ones, using TiMBL's confidence score.
7. WEAK MATCH. A simplified version of STRONG MATCH: not all mentions in the growing entity need to be classified as coreferent with the mention under analysis. A single positive pairwise decision suffices for the mention to be clustered into that entity.[10]

[7] The opposite search direction was also tried but gave worse results.

[8] Taking the first mention classified as coreferent follows SOON *et al.* (2001)'s first-link approach.

[9] In TiMBL, being a memory-based learner, the closer the distance to an instance, the more confident the decision. We chose 0.09 because it appeared to offer the best results.

[10] STRONG and WEAK MATCH are similar to LUO *et al.* (2004)'s entity-mention and mention-pair models.

4.3.3 Features

We follow SOON *et al.* (2001), NG i CARDIE (2002*b*) and LUO *et al.* (2004) to generate most of the 29 features we use for the pairwise model. These include features that capture information from different linguistic levels: textual strings (head match, substring match, distance, frequency), morphology (mention type, coordination, possessive phrase, gender match, number match), syntax (nominal predicate, apposition, relative clause, grammatical function), and semantic match (named-entity type, is-a type, supertype).

For Spanish, we use 34 features as a few variations are needed for language-specific issues such as zero subjects (RECASENS i HOVY, 2009).

4.3.4 Evaluation

Since they sometimes provide quite different results, we evaluate using three coreference measures, as there is no agreement on a standard.

- MUC (VILAIN *et al.*, 1995). It computes the number of links common between the true and system partitions. Recall (R) and precision (P) result from dividing it by the minimum number of links required to specify the true and the system partitions, respectively.
- B^3 (BAGGA i BALDWIN, 1998). R and P are computed for each mention and averaged at the end. For each mention, the number of common mentions between the true and the system entity is divided by the number of mentions in the true entity or in the system entity to obtain R and P, respectively.
- CEAF (LUO, 2005). It finds the best one-to-one alignment between true and system entities. Using true mentions and the ϕ_3 similarity function, R and P are the same and correspond to the number of common mentions between the aligned entities divided by the total number of mentions.

4.4 Parameter 1: Language

The first experiment compared the performance of a coreference resolution system on a Germanic and a Romance language—English and Spanish—to explore to what extent language-specific issues such as zero subjects[11] or grammatical gender might influence a system.

Although OntoNotes and AnCora are two different corpora, they are very similar in those aspects that matter most for the study's purpose: they both include a substantial amount of texts belonging to the same genre (news) and manually annotated from the morphological to the semantic levels (POS tags, syntactic constituents, NEs, WordNet synsets, and coreference relations). More importantly,

[11] Most Romance languages are pro-drop allowing zero subject pronouns, which can be inferred from the verb.

		#docs	#words	#mentions	#entities	#singleton entities	#multi-mention entities
AnCora	Train	955	299,014	91,904	64,535	54,991	9,544
	Test	30	9,851	2,991	2,189	1,877	312
OntoNotes	Train	850	301,311	74,692	55,819	48,199	7,620
	Test	33	9,763	2,463	1,790	1,476	314

Table 4.1: Corpus statistics for the large portion of OntoNotes and AnCora

	AnCora	OntoNotes
Pronouns	14.09	17.62
Personal pronouns	2.00	12.10
Zero subject pronouns	6.51	–
Possessive pronouns	3.57	2.96
Demonstrative pronouns	0.39	1.83
Definite NPs	37.69	20.67
Indefinite NPs	7.17	8.44
Demonstrative NPs	1.98	3.41
Bare NPs	33.02	42.92
Misc.	6.05	6.94

Table 4.2: Mention types (%) in Table 4.1 datasets

very similar coreference annotation guidelines make AnCora the ideal Spanish counterpart to OntoNotes.

Datasets Two datasets of similar size were selected from AnCora and OntoNotes in order to rule out corpus size as an explanation of any difference in performance. Corpus statistics about the distribution of mentions and entities are shown in Tables 4.1 and 4.2. Given that this paper is focused on coreference between NPs, the number of mentions only includes NPs. Both AnCora and OntoNotes annotate only multi-mention entities (i.e., those containing two or more coreferent mentions), so singleton entities are assumed to correspond to NPs with no coreference annotation.

Apart from a larger number of mentions in Spanish (Table 4.1), the two datasets look very similar in the distribution of singletons and multi-mention entities: about 85% and 15%, respectively. Multi-mention entities have an average of 3.9 mentions per entity in AnCora and 3.5 in OntoNotes. The distribution of mention types (Table 4.2), however, differs in two important respects: AnCora has a smaller number of personal pronouns as Spanish typically uses zero subjects, and it has a smaller number of bare NPs as the definite article accompanies more NPs than in English.

	MUC			B^3			CEAF
	P	R	F	P	R	F	P / R / F
AnCora - Spanish							
1. ALL SINGLETONS	–	–	–	100	73.32	84.61	73.32
2. HEAD MATCH	55.03	37.72	44.76	91.12	79.88	**85.13**	75.96
3. HEAD MATCH + PRON	48.22	44.24	46.14	86.21	80.66	83.34	76.30
4. STRONG MATCH	45.64	51.88	48.56	80.13	82.28	81.19	75.79
5. SUPER STRONG MATCH	45.68	36.47	40.56	86.10	79.09	82.45	**77.20**
6. BEST MATCH	43.10	35.59	38.98	85.24	79.67	82.36	75.23
7. WEAK MATCH	45.73	65.16	**53.75**	68.50	87.71	76.93	69.21
OntoNotes - English							
1. ALL SINGLETONS	–	–	–	100	72.68	84.18	72.68
2. HEAD MATCH	55.14	39.08	45.74	90.65	80.87	**85.48**	76.05
3. HEAD MATCH + PRON	47.10	53.05	49.90	82.28	83.13	82.70	75.15
4. STRONG MATCH	47.94	55.42	51.41	81.13	84.30	82.68	78.03
5. SUPER STRONG MATCH	48.27	47.55	47.90	84.00	82.27	83.13	78.24
6. BEST MATCH	50.97	46.66	48.72	86.19	82.70	84.41	**78.44**
7. WEAK MATCH	47.46	66.72	**55.47**	70.36	88.05	78.22	71.21

Table 4.3: CISTELL results varying the corpus language

Results and discussion Table 4.3 presents CISTELL's results for each dataset. They make evident problems with the evaluation metrics, namely the fact that the generated rankings are contradictory (DENIS i BALDRIDGE, 2009). They are consistent across the two corpora though: MUC rewards WEAK MATCH the most, B^3 rewards HEAD MATCH the most, and CEAF is divided between SUPER STRONG MATCH and BEST MATCH.

These preferences seem to reveal weaknesses of the scoring methods that make them biased towards a type of output. The model preferred by MUC is one that clusters many mentions together, thus getting a large number of correct coreference links (notice the high R for WEAK MATCH), but also many spurious links that are not duly penalized. The resulting output is not very desirable.[12] In contrast, B^3 is more P-oriented and scores conservative outputs like HEAD MATCH and BEST MATCH first, even if R is low. CEAF achieves a better compromise between P and R, as corroborated by the quality of the output.

The baselines and the system runs perform very similarly in the two corpora, but slightly better for English. It seems that language-specific issues do not result in significant differences—at least for English and Spanish—once the feature set has been appropriately adapted, e.g., including features about zero subjects or removing those about possessive phrases. Comparing the feature ranks, we find that the features that work best for each language largely overlap and are language independent, like head match, is-a match, and whether the mentions are pronominal.

[12] Due to space constraints, the actual output cannot be shown here. We are happy to send it to interested requesters.

		#docs	#words	#mentions	#entities	#singleton entities	#multi-mention entities
OntoNotes	Train	297	87,068	22,127	15,983	13,587	2,396
	Test	33	9,763	2,463	1,790	1,476	314
ACE	Train	297	87,068	12,951	5,873	3,599	2,274
	Test	33	9,763	1,464	746	459	287

Table 4.4: Corpus statistics for the aligned portion of ACE and OntoNotes on gold-standard data

4.5 Parameter 2: Annotation scheme

In the second experiment, we used the 100k-word portion (from the TDT collection) shared by the OntoNotes and ACE corpora (330 OntoNotes documents occurred as 22 ACE-2003 documents, 185 ACE-2004 documents, and 123 ACE-2005 documents). CISTELL was trained on the same texts in both corpora and applied to the remainder. The three measures were then applied to each result.

Datasets Since the two annotation schemes differ significantly, we made the results comparable by mapping the ACE entities (the simpler scheme) onto the information contained in OntoNotes.[13] The mapping allowed us to focus exclusively on the differences expressed on both corpora: the types of mentions that were annotated, the definition of identity of reference, etc.

Table 4.4 presents the statistics for the OntoNotes dataset merged with the ACE entities. The mapping was not straightforward due to several problems: there was no match for some mentions due to syntactic or spelling reasons (e.g., *El Popo* in OntoNotes vs. *Ell Popo* in ACE). ACE mentions for which there was no parse tree node in the OntoNotes gold-standard tree were omitted, as creating a new node could have damaged the tree.

Given that only seven entity types are annotated in ACE, the number of OntoNotes mentions is almost twice as large as the number of ACE mentions. Unlike OntoNotes, ACE mentions include premodifiers (e.g., *state* in *state lines*), national adjectives (e.g., *Iraqi*) and relative pronouns (e.g., *who, that*). Also, given that ACE entities correspond to types that are usually coreferred (e.g., people, organizations, etc.), singletons only represent 61% of all entities, while they are 85% in OntoNotes. The average entity size is 4 in ACE and 3.5 in OntoNotes.

A second major difference is the definition of coreference relations, illustrated here:

(2) [This] was [an all-white, all-Christian community that all the sudden was taken over ... by different groups].

(3) [[Mayor] John Hyman] has a simple answer.

[13] Both ACE entities and types were mapped onto the OntoNotes dataset.

	MUC			B^3			CEAF
	P	R	F	P	R	F	P / R / F
OntoNotes scheme							
1. ALL SINGLETONS	–	–	–	100	72.68	84.18	72.68
2. HEAD MATCH	55.14	39.08	45.74	90.65	80.87	**85.48**	76.05
3. HEAD MATCH + PRON	47.10	53.05	49.90	82.28	83.13	82.70	75.15
4. STRONG MATCH	46.81	53.34	49.86	80.47	83.54	81.97	76.78
5. SUPER STRONG MATCH	46.51	40.56	43.33	84.95	80.16	82.48	76.70
6. BEST MATCH	52.47	47.40	49.80	86.10	82.80	84.42	**77.87**
7. WEAK MATCH	47.91	64.64	**55.03**	71.73	87.46	78.82	71.74
ACE scheme							
1. ALL SINGLETONS	–	–	–	100	50.96	67.51	50.96
2. HEAD MATCH	82.35	39.00	52.93	95.27	64.05	**76.60**	66.46
3. HEAD MATCH + PRON	70.11	53.90	60.94	86.49	68.20	76.27	68.44
4. STRONG MATCH	64.21	64.21	64.21	76.92	73.54	75.19	**70.01**
5. SUPER STRONG MATCH	60.51	56.55	58.46	76.71	69.19	72.76	66.87
6. BEST MATCH	67.50	56.69	61.62	82.18	71.67	76.57	69.88
7. WEAK MATCH	63.52	80.50	**71.01**	59.76	86.36	70.64	64.21

Table 4.5: CISTELL results varying the annotation scheme on gold-standard data

(4) [Postville] now has 22 different nationalities ... For those who prefer [the old Postville], Mayor John Hyman has a simple answer.

In ACE, nominal predicates corefer with their subject (2), and appositive phrases corefer with the noun they are modifying (3). In contrast, they do not fall under the identity relation in OntoNotes, which follows the linguistic understanding of coreference according to which nominal predicates and appositives express properties of an entity rather than refer to a second (coreferent) entity (VAN DEEMTER i KIBBLE, 2000). Finally, the two schemes frequently disagree on borderline cases in which coreference turns out to be especially complex (4). As a result, some features will behave differently, e.g., the appositive feature has the opposite effect in the two datasets.

Results and discussion From the differences pointed out above, the results shown in Table 4.5 might be surprising at first. Given that OntoNotes is not restricted to any semantic type and is based on a more sophisticated definition of coreference, one would not expect a system to perform better on it than on ACE. The explanation is given by the ALL SINGLETONS baseline, which is 73–84% for OntoNotes and only 51–68% for ACE. The fact that OntoNotes contains a much larger number of singletons—as Table 4.4 shows—results in an initial boost of performance (except with the MUC score, which ignores singletons). In contrast, the score improvement achieved by HEAD MATCH is much more noticeable on ACE than on OntoNotes, which indicates that many of its coreferent mentions share the same head.

The systematic biases of the measures that were observed in Table 4.3 appear

		#docs	#words	#mentions	#entities	#singleton entities	#multi-mention entities
OntoNotes	Train	297	80,843	16,945	12,127	10,253	1,874
	Test	33	9,073	1,931	1,403	1,156	247
ACE	Train	297	80,843	13,648	6,041	3,652	2,389
	Test	33	9,073	1,537	775	475	300

Table 4.6: Corpus statistics for the aligned portion of ACE and OntoNotes on automatically parsed data

again in the case of MUC and B^3. CEAF is divided between BEST MATCH and STRONG MATCH. The higher value of the MUC score for ACE is another indication of its tendency to reward correct links much more than to penalize spurious ones (ACE has a larger proportion of multi-mention entities).

The feature rankings obtained for each dataset generally coincide as to which features are ranked best (namely NE match, is-a match, and head match), but differ in their particular ordering.

It is also possible to compare the OntoNotes results in Tables 4.3 and 4.5, the only difference being that the first training set was three times larger. Contrary to expectation, the model trained on a larger dataset performs just slightly better. The fact that more training data does not necessarily lead to an increase in performance conforms to the observation that there appear to be few general rules (e.g., head match) that systematically govern coreference relationships; rather, coreference appeals to individual unique phenomena appearing in each context, and thus after a point adding more training data does not add much new generalizable information. Pragmatic information (discourse structure, world knowledge, etc.) is probably the key, if ever there is a way to encode it.

4.6 Parameter 3: Preprocessing

The goal of the third experiment was to determine how much the source and nature of preprocessing information matters. Since it is often stated that coreference resolution depends on many levels of analysis, we again compared the two corpora, which differ in the amount and correctness of such information. However, in this experiment, entity mapping was applied in the opposite direction: the OntoNotes entities were mapped onto the automatically preprocessed ACE dataset. This exposes the shortcomings of automated preprocessing in ACE for identifying all the mentions identified and linked in OntoNotes.

Datasets The ACE data was morphologically annotated with a tokenizer based on manual rules adapted from the one used in CoNLL (TJONG KIM SANG i DE MEULDER, 2003), with TnT 2.2, a trigram POS tagger based on Markov models (BRANTS, 2000), and with the built-in WordNet lemmatizer (FELLBAUM,

	MUC			B^3			CEAF
	P	R	F	P	R	F	P / R / F
OntoNotes scheme							
1. ALL SINGLETONS	–	–	–	100	72.66	84.16	72.66
2. HEAD MATCH	56.76	35.80	43.90	92.18	80.52	**85.95**	76.33
3. HEAD MATCH + PRON	47.44	54.36	50.66	82.08	83.61	82.84	74.83
4. STRONG MATCH	52.66	58.14	55.27	83.11	85.05	84.07	**78.30**
5. SUPER STRONG MATCH	51.67	46.78	49.11	85.74	82.07	83.86	77.67
6. BEST MATCH	54.38	51.70	53.01	86.00	83.60	84.78	78.15
7. WEAK MATCH	49.78	64.58	**56.22**	75.63	87.79	81.26	74.62
ACE scheme							
1. ALL SINGLETONS	–	–	–	100	50.42	67.04	50.42
2. HEAD MATCH	81.25	39.24	52.92	94.73	63.82	**76.26**	65.97
3. HEAD MATCH + PRON	69.76	53.28	60.42	86.39	67.73	75.93	**68.05**
4. STRONG MATCH	58.85	58.92	58.89	73.36	70.35	71.82	66.30
5. SUPER STRONG MATCH	56.19	50.66	53.28	75.54	66.47	70.72	63.96
6. BEST MATCH	63.38	49.74	55.74	80.97	68.11	73.99	65.97
7. WEAK MATCH	60.22	78.48	**68.15**	55.17	84.86	66.87	59.08

Table 4.7: CISTELL results varying the annotation scheme on automatically preprocessed data

1998). Syntactic chunks were obtained from YamCha 1.33, an SVM-based NP-chunker (KUDOH i MATSUMOTO, 2000), and parse trees from Malt Parser 0.4, an SVM-based parser (HALL *et al.*, 2007).

Although the number of words in Tables 4.4 and 4.6 should in principle be the same, the latter contains fewer words as it lacks the null elements (traces, elipsed material, etc.) manually annotated in OntoNotes. Missing parse tree nodes in the automatically parsed data account for the considerably lower number of OntoNotes mentions (approx. 5,700 fewer mentions).[14] However, the proportions of singleton:multi-mention entities as well as the average entity size do not vary.

Results and discussion The ACE scores for the automatically preprocessed models in Table 4.7 are about 3% lower than those based on OntoNotes gold-standard data in Table 4.5, providing evidence for the advantage offered by gold-standard preprocessing information. In contrast, the similar—if not higher—scores of OntoNotes can be attributed to the use of the annotated ACE entity types. The fact that these are annotated not only for proper nouns (as predicted by an automatic NER) but also for pronouns and full NPs is a very helpful feature for a coreference resolution system.

Again, the scoring metrics exhibit similar biases, but note that CEAF prefers HEAD MATCH + PRON in the case of ACE, which is indicative of the noise brought by automatic preprocessing.

[14] In order to make the set of mentions as similar as possible to the set in Section 4.5, OntoNotes singletons were mapped from the ones detected in the gold-standard treebank.

A further insight is offered from comparing the feature rankings with gold-standard syntax to that with automatic preprocessing. Since we are evaluating now on the ACE data, the NE match feature is also ranked first for OntoNotes. Head and is-a match are still ranked among the best, yet syntactic features are not. Instead, features like NP type have moved further up. This reranking probably indicates that if there is noise in the syntactic information due to automatic tools, then morphological and syntactic features switch their positions.

Given that the noise brought by automatic preprocessing can be harmful, we tried leaving out the grammatical function feature. Indeed, the results increased about 2–3%, STRONG MATCH scoring the highest. This points out that conclusions drawn from automatically preprocessed data about the kind of knowledge relevant for coreference resolution might be mistaken. Using the most successful basic features can lead to the best results when only automatic preprocessing is available.

4.7 Conclusion

Regarding evaluation, the results clearly expose the systematic tendencies of the evaluation measures. The way each measure is computed makes it biased towards a specific model: MUC is generally too lenient with spurious links, B^3 scores too high in the presence of a large number of singletons, and CEAF does not agree with either of them. It is a cause for concern that they provide contradictory indications about the core of coreference, namely the resolution models—for example, the model ranked highest by B^3 in Table 4.7 is ranked *lowest* by MUC. We always assume evaluation measures provide a 'true' reflection of our approximation to a gold standard in order to guide research in system development and tuning.

Further support to our claims comes from the results of SemEval-2010 Task 1 (RECASENS *et al.*, 2010*b*). The performance of the six participating systems shows similar problems with the evaluation metrics, and the singleton baseline was hard to beat even by the highest-performing systems.

Since the measures imply different conclusions about the nature of the corpora and the preprocessing information applied, should we use them now to constrain the ways our corpora are created in the first place, and what preprocessing we include or omit? Doing so would seem like circular reasoning: it invalidates the notion of the existence of a true and independent gold standard. But if apparently incidental aspects of the corpora can have such effects—effects rated quite differently by the various measures—then we have no fixed ground to stand on.

The worrisome fact that there is currently no clearly preferred and 'correct' evaluation measure for coreference resolution means that we cannot draw definite conclusions about coreference resolution systems at this time, unless they are compared on exactly the same corpus, preprocessed under the same conditions, and all three measures agree in their rankings.

Acknowledgments We thank Dr. M. Antònia Martí for her generosity in allowing the

first author to visit ISI to work with the second. Special thanks to Edgar Gonzàlez for his kind help with conversion issues. This work was partially supported by the Spanish Ministry of Education through an FPU scholarship (AP2006-00994) and the TEXT-MESS 2.0 Project (TIN2009-13391-C04-04).

CAPÍTOL 5

BLANC: Implementing the Rand Index for Coreference Evaluation

Marta Recasens* and Eduard Hovy**

*University of Barcelona
**USC Information Sciences Institute

To appear in *Natural Language Engineering*

Abstract This article addresses the current state of coreference resolution evaluation, in which different measures (notably, MUC, B^3, CEAF, and ACE-value) are applied in different studies. None of them is fully adequate, and their measures are not commensurate. We enumerate the desiderata for a coreference scoring measure, discuss the strong and weak points of the existing measures, and propose the BiLateral Assessment of Noun-phrase Coreference, a variation of the Rand index created to suit the coreference task. The BiLateral Assessment of Noun-phrase Coreference rewards both coreference and non-coreference links by averaging the F-scores of the two types, does not ignore singletons—the main problem with the MUC score—and does not inflate the score in their presence—a problem with the B^3 and CEAF scores. In addition, its fine granularity is consistent over the whole range of scores and affords better discrimination between systems.

5.1 Introduction

Coreference resolution is the task of determining which expressions in a text refer to the same entity or event. At heart, the problem is one of grouping into 'equivalence classes' all mentions that corefer and none that do not, which is a kind of

clustering. But since documents usually contain many referring expressions, many different combinations are possible, and measuring partial cluster correctness, especially since *sameness* is transitive, makes evaluation difficult. One has to assign scores to configurations of correct and incorrect links in a way that reflects intuition and is consistent. Different assignment policies have resulted in different evaluation measures that deliver quite different patterns of scores. Among the different scoring measures that have been developed, four are generally used: MUC (VILAIN *et al.*, 1995), B^3 (BAGGA i BALDWIN, 1998), CEAF (LUO, 2005), and the ACE-value (DODDINGTON *et al.*, 2004).

Unfortunately, despite the measures being incommensurate, researchers often use only one or two measures when evaluating their systems. For example, some people employ the (older) MUC measure in order to compare their results with previous work (HAGHIGHI i KLEIN, 2007; YANG *et al.*, 2008); others adopt the more recent advances and use either B^3, CEAF, or the ACE-value (CULOTTA *et al.*, 2007; DAUMÉ III i MARCU, 2005); and a third group includes two or more scores for the sake of completeness (LUO *et al.*, 2004; BENGTSON i ROTH, 2008; NG, 2009; FINKEL i MANNING, 2008; POON i DOMINGOS, 2008).

This situation makes it hard to successfully compare systems, hindering the progress of research in coreference resolution. There is a pressing need to (1) define what exactly a scoring metric for coreference resolution needs to measure; (2) understand the advantages and disadvantages of each of the existing measures; and (3) reach agreement on a standard measure(s). This article addresses the first two questions—we enumerate the desiderata for an adequate coreference scoring measure, and we compare the different existing measures—and proposes the BiLateral Assessment of Noun-phrase Coreference (BLANC) measure. BLANC adapts the Rand index RAND (1971) to coreference addressing observed shortcomings in a simple fashion to obtain a fine granularity that allows better discrimination between systems.

The article is structured as follows. Section 5.2 considers the difficulties of evaluating coreference resolution. Section 5.3 gives an overview of the existing measures, highlighting their advantages and drawbacks, and lists some desiderata for an ideal measure. In Section 5.4, the BLANC measure is presented in detail. Section 5.5 shows the discriminative power of BLANC by comparing its scores to those of the other measures on artificial and real data, and provides illustrative plots. Finally, conclusions are drawn in Section 5.6.

5.2 Coreference resolution and its evaluation: an example

Coreference resolution systems assign each mention (usually a noun phrase) in the text to the entity it refers to and thereby link coreferent mentions into chains.[1]

[1]Following the terminology of the Automatic Content Extraction (ACE) program, a **mention** is defined as an instance of reference to an object, and an **entity** is the collection of mentions referring

[Eyewitnesses]$_{m_1}$ reported that [Palestinians]$_{m_2}$ demonstrated today Sunday in [the West Bank]$_{m_3}$ against [the [Sharm el-Sheikh]$_{m_4}$ summit to be held in [Egypt]$_{m_6}$]$_{m_5}$. In [Ramallah]$_{m_7}$, [around 500 people]$_{m_8}$ took to [[the town]$_{m_9}$'s streets]$_{m_{10}}$ chanting [slogans]$_{m_{11}}$ denouncing [the summit]$_{m_{12}}$ and calling on [Palestinian leader Yasser Arafat]$_{m_{13}}$ not to take part in [it]$_{m_{14}}$.

Figure 5.1: Example of coreference (from ACE-2004)

Some entities are expressed only once (singletons), whereas others are referred multiple times (multi-mention entities). Only multi-mention entities contain coreferent mentions. For example, in the text segment of Fig. 5.1, we find the following:

- Nine singletons: {*eyewitnesses*}$_{\text{G1}}$, {*Palestinians*}$_{\text{G2}}$, {*the West Bank*}$_{\text{G3}}$, {*Sharm el-Sheikh*}$_{\text{G4}}$, {*Egypt*}$_{\text{G5}}$, {*around 500 people*}$_{\text{G6}}$, {*the town's streets*}$_{\text{G7}}$, {*slogans*}$_{\text{G8}}$, {*Palestinian leader Yasser Arafat*}$_{\text{G9}}$
- One two-mention entity: {*Ramallah*, *the town*}$_{\text{G10}}$
- One three-mention entity: {*the Sharm el-Sheikh summit to be held in Egypt*, *the summit*, *it*}$_{\text{G11}}$

In evaluating the output produced by a coreference resolution system, we need to compare the true set of entities (the **gold partition**, GOLD, produced by human expert) with the predicted set of entities (the **system partition**, SYS, produced by the system or human to be evaluated). The mentions in GOLD are known as **true mentions**, and the mentions in SYS are known as **system mentions**. Let a system produce the following partition for the same example in Fig. 5.1:

- Seven singletons: {*eyewitnesses*}$_{\text{S1}}$, {*Palestinians*}$_{\text{S2}}$, {*the West Bank*}$_{\text{S3}}$, {*around 500 people*}$_{\text{S4}}$, {*the town's streets*}$_{\text{S5}}$, {*slogans*}$_{\text{S6}}$, {*Palestinian leader Yasser Arafat*}$_{\text{S7}}$
- Two two-mention entities: {*Sharm el-Sheikh*, *Egypt*}$_{\text{S8}}$, {*the Sharm el-Sheikh summit to be held in Egypt*, *the summit*}$_{\text{S9}}$
- One three-mention entity: {*Ramallah*, *the town*, *it*}$_{\text{S10}}$

Schematically, the comparison problem is illustrated in Fig. 5.2. Some links are missed and others are wrongly predicted; e.g., entity S9 is missing one mention (compare with G11), whereas S10 includes a wrong mention, and two non-coreferent mentions are linked under S8. The difficulty of evaluating coreference resolution arises from the interaction of the issues that have to be addressed simultaneously: Should we focus on the number of correct coreference links? Or should we instead take each equivalence class as the unit of evaluation? Do we reward singletons with the same weight that we reward a multi-mention entity? Different

to the same object in a document.

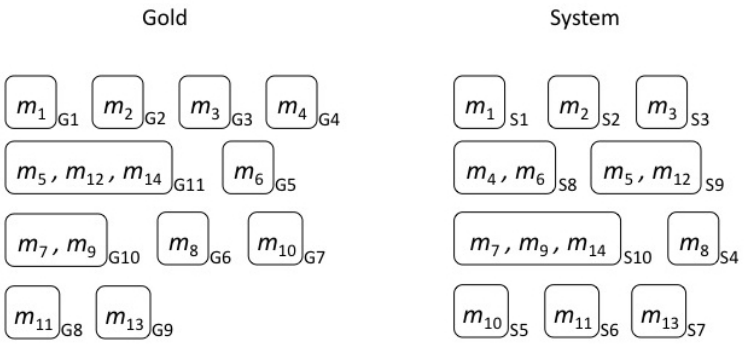

Figure 5.2: The problem of comparing the gold partition with the system partition for a given text (Fig. 5.1)

decisions will result in different evaluation scores, which will determine how good SYS is considered to be in comparison with GOLD.

The evaluation measures developed to date all make somewhat different decisions on these points. While these decisions have been motivated in terms of one or another criterion, they also have unintended unsatisfactory consequences. We next review some current measures and identify the desiderata for a coreference measure.

5.3 Current measures and desiderata for the future

5.3.1 Current measures: strong and weak points

This section reviews the main advantages and drawbacks of the principal coreference evaluation measures. The main difference resides in the way they conceptualize how a coreference set within a text is defined: either in terms of **links**, i.e., the pairwise links between mentions (MUC, Pairwise F1, Rand), or in terms of **classes** or **clusters**, i.e., the entities (B^3, CEAF, ACE-value, mutual information). Although the two approaches are equivalent in that knowing the links allows building the coreference classes, and knowing the classes allows inferring the links, differences in instantiation design produce a range of evaluation metrics that vary to such an extent that still today there is no widely agreed upon standard. Table 5.1 shows how the different system outputs in Fig. 5.3 (borrowed from LUO (2005)) are scored by the various scoring algorithms presented next.

MUC (VILAIN *et al.*, 1995). This is the oldest and most widely used measure, defined as part of the MUC-6 and MUC-7 evaluation tasks on coreference resolution. It relies on the notion that the minimum number of links needed to specify either GOLD or SYS is the total number of mentions minus the number of entities. The MUC measure computes the number of all coreference links common between

System response	MUC-F	B^3-F	CEAF	F1	H	Rand
(a)	94.7	86.5	83.3	80.8	77.8	84.8
(b)	94.7	73.7	58.3	63.6	57.1	62.1
(c)	90.0	54.5	41.7	48.3	0	31.8
(d)	—	40.0	25.0	—	48.7	68.2

Table 5.1: Comparison of evaluation metrics on the examples in Fig. 5.3

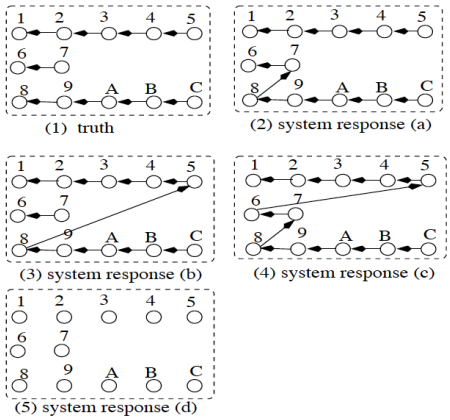

Figure 5.3: Example entity partitions (from LUO (2005))

GOLD and SYS. To obtain recall (R), this number is divided by the minimum number of links required to specify GOLD. To obtain precision (P), it is divided by the minimum number of links required to specify SYS.

As observed by BAGGA i BALDWIN (1998) and LUO (2005), the MUC metric is severely flawed for two main reasons. First, it is indulgent as it is based on the *minimal* number of missing and wrong links, which often results in counterintuitive results. Classifying one mention into a wrong entity counts as one P and one R error, while completely merging two entities counts as a single R error, although this is further away from the real answer. As a result, the MUC score is too lenient with systems that produce overmerged entities (entity sets containing many referring expressions), as shown by system responses (b) and (c) in Table 5.1. If all mentions in each document of the MUC test sets[2] are linked into one single entity, the MUC metric gives a score higher than any published system (FINKEL i MANNING, 2008). Second, given that it only takes into account coreference links, the addition of singletons to SYS does not make any difference. It is only when a singleton mention is misclassified in a multi-mention entity that the MUC score decreases. This is why the entry for system response (d) in Table 5.1 is empty.

[2]The MUC-6 and MUC-7 corpora were only annotated with multi-mention entities (HIRSCHMAN i CHINCHOR, 1997).

	ACE-2004 (English)		AnCora-Es (Spanish)	
	#	%	#	%
Mentions	28,880	100.00	88,875	100.00
Entities	11,989	100.00	64,421	100.00
Singletons	7,305	60.93	55,264	85.79
2-mention	2,126	17.73	4,825	7.49
3-mention	858	7.16	1,711	2.66
4-mention	479	4.00	869	1.35
5-mention	287	2.39	485	0.75
6-10-mention	567	4.73	903	1.40
$>$ 11-mention	367	3.06	364	0.57

Table 5.2: Distribution of mentions into entities in two corpora: the English ACE-2004 and the Spanish AnCora-Es

B^3 (BAGGA i BALDWIN, 1998). To penalize clustering too many mentions in the same entity, this metric computes R and P for each mention, including singletons. The total number of intersecting mentions between the GOLD and SYS entities is computed and divided by the total number of mentions in the GOLD entity to obtain R, or in the SYS entity to obtain P. The average over the individual mention scores gives the final scores.

Although B^3 addresses the shortcomings of MUC, it presents a drawback in that scores squeeze up too high due to singletons: when many singletons are present, scores rapidly approach 100%. This leaves little numerical room for comparing systems, and forces one to consider differences in the second and third decimal places when scores are high (while such differences are meaninglessly small in lower ranges). It is not possible to observe this in Table 5.1 as the truth in Fig. 5.3 does not contain any singleton. However, it turns out that singletons are the largest group in real texts (see Table 5.2): about 86% of the entities if the entire set of mentions is considered, like in the AnCora corpora; 61% of the entities in the ACE corpora, where the coreference annotation is restricted to seven semantic types (person, organization, geo-political entity, location, facility, vehicle, and weapon). A side effect is that B^3 scores are inflated, obscuring the intuitively appropriate level of accuracy of a system in terms of coreference links.

CEAF (LUO, 2005). LUO (2005) considers that B^3 can give counterintuitive results due to the fact that an entity can be used more than once when aligning the entities in GOLD and SYS. In Fig. 5.3, B^3-R is 100% for system response (c) even though the true set of entities has not been found; conversely, B^3-P is 100% for system response (d) even though not all the SYS entities are correct. Thus, he proposes CEAF, which finds the best one-to-one mapping between the entities in GOLD and SYS, i.e., each SYS entity is aligned with at most one GOLD entity, and the best alignment is the one maximizing the similarity. Depending on the

similarity function, LUO (2005) distinguishes between the mention-based CEAF and the entity-based CEAF, but we will focus on the former as it is the most widely used. It employs Luo's (2005) ϕ_3 similarity function. When true mentions are used, R and P scores are the same. They correspond to the number of common mentions between every two aligned entities divided by the total number of mentions.

CEAF, however, suffers from the singleton problem just as B^3 does. This accounts for the fact that the B^3 and CEAF scores are usually higher than the MUC on corpora where singletons are annotated (e.g., ACE, AnCora), because a great percentage of the score is simply due to the resolution of singletons. In addition, CEAF's entity alignment might cause a correct coreference link to be ignored if that entity finds no alignment in GOLD (DENIS i BALDRIDGE, 2009). Finally, all entities are weighted equally, irrespective of the number of mentions they contain (STOYANOV *et al.*, 2009), so that creating a wrong entity composed of two small entities is penalized to the same degree as creating a wrong entity composed of a small and a large entity.

ACE-value (DODDINGTON *et al.*, 2004). The ACE-value, the official metric in the ACE program, is very task-specific, and not really useful for the general coreference problem that is not limited to a set of specific semantic types. A score is computed by subtracting a normalized cost from 1. The normalized cost corresponds to the sum of errors produced by unmapped and missing mentions/entities as well as wrong mentions/entities,[3] normalized against the cost of a system that does not output any entity. Each error has an associated cost that depends on the type of ACE-entity and on the kind of mention, but these costs have changed between successive evaluations. The ACE-value is hard to interpret (LUO, 2005): a system with 90% does not mean that 90% of system entities or mentions are correct, but that the cost of the system, relative to the one producing no entity, is 10%.

Pairwise F1. Also known as positive-link-identification F-score. If reported, this metric is always included in addition to MUC, B^3 and/or CEAF, as it is meant to give some further insight not provided by the other metrics (CHOI i CARDIE, 2007; POON i DOMINGOS, 2008; HAGHIGHI i KLEIN, 2009). Pairwise F1 simply computes P, R, and F over all pairs of coreferent mentions. As noted by HAGHIGHI i KLEIN (2009), merging or separating entities is over-penalized quadratically in the number of mentions. Besides, it ignores the correct identification of singletons.

Mutual information, H (POPESCU-BELIS, 2000). The H measure draws on information theory to evaluate coreference resolution. GOLD and SYS are seen as the two ends of the communication channel, GOLD being the sender or speaker, and SYS being the receiver or the hearer. The coreference information of GOLD

[3] In the ACE evaluation program, mentions and entities in SYS that are not mapped onto any mention or entity in GOLD receive a false alarm penalty.

and SYS correspond to the entropy of GOLD and SYS, respectively. Then the GOLD and SYS partitions are compared on the basis of mutual coreference information. R is obtained by subtracting the conditioned entropy of GOLD given SYS (loss of information) from the entropy of GOLD. P is obtained by subtracting the conditioned entropy of SYS given GOLD (irrelevant information gains) from the entropy of SYS. Both values are then normalized. This measure has been hardly used for reporting results of real systems, and it emerges from the results reported by POPESCU-BELIS (2000) that H is not superior to the other existing measures. Popescu-Belis concludes that each metric, by focusing on different aspects of the data, provides a different perspective on the quality of the system answer.

Rand index (RAND, 1971). The Rand index is a general clustering evaluation metric that measures the similarity between two clusterings (i.e., partitions) by considering how each pair of data points is assigned in each clustering. Stated in coreference terms, the Rand index equals the number of mention pairs that are either placed in an entity or assigned to separate entities in both GOLD and SYS, normalized by the total number of mention pairs in each partition. The motivations behind this measure are three (where we replace 'point' by 'mention', 'cluster' by 'entity', and 'clustering' by 'entity partition'): (1) every mention is unequivocally assigned to a specific entity; (2) entities are defined just as much by those points which they do not contain as by those mentions which they do contain; and (3) all mentions are of equal importance in the determination of the entity partition.

The only use of the Rand index for coreference resolution appears in FINKEL i MANNING (2008). Although Rand has the potential to capture well the coreference problem, it is not useful if applied as originally defined due to the significant imbalance between the number of coreferent mentions and the number of singletons (Table 5.2). The extremely high number of mention pairs that are found in different entities in GOLD and SYS explains the high figures obtained by all systems reported in FINKEL i MANNING (2008), and by system response (d) in Table 5.1. Hence, the low discriminatory power of Rand. The BLANC measure that we introduce in Section 5.4 implements Rand in a way suited to the coreference problem.

It is often hard for researchers working on coreference resolution to make sense of the state of the art. Compare, for example, the scores shown in Table 5.3 that correspond to various systems[4] and two baselines: (1) all singletons (i.e., no coreference link is created, but each mention is considered to be a separate entity), and (2) one entity (i.e., all document mentions are clustered into one single entity). The only measure for which we have the results of all systems is MUC, but this is the one with the largest number of drawbacks, as evidenced by the high score of the one-entity baseline. It is clear that the measures do not produce the same ranking of the systems, other than the fact that they all rank LUO *et al.* (2004) and LUO

[4]Scores published here but missing in the original papers were computed by us from the authors' outputs.

System	MUC-F	B^3-F	CEAF	ACE-value
		ACE-2		
All-singletons baseline	—	55.9	38.8	
One-entity baseline	76.5	17.3	21.7	
LUO *et al.* (2004)	80.7	77.0	73.2	89.8
FINKEL i MANNING (2008)	64.1	73.8		
POON i DOMINGOS (2008)	68.4	69.2	63.9	
DENIS i BALDRIDGE (2009)	70.1	72.7	66.2	
NG (2009)	61.3		61.6	
		ACE-2004		
All-singletons baseline	—	59.0	41.8	
One-entity baseline	74.4	17.8	21.4	
LUO i ZITOUNI (2005)	86.0	83.7	82.0	91.6
HAGHIGHI i KLEIN (2007)	63.3			
BENGTSON i ROTH (2008)	75.8	80.8	75.0	
POON i DOMINGOS (2008)	69.1	71.2	65.9	
WICK i MCCALLUM (2009)	70.1	81.5		

Table 5.3: Performance of state-of-the-art coreference systems on ACE

i ZITOUNI (2005) as the best systems for each data set. This sort of discrepancy makes it impossible in the long term to conduct research on this question: which measure should one trust, and why?

Apart from the pros and cons of each measure, the difficulty in comparing the performance of different coreference resolution systems is compounded by other factors, such as the use of true or system mentions and the use of different test sets (STOYANOV *et al.*, 2009). Some systems in Table 5.3 are not directly comparable since testing on a different set of mentions or on a different data set is likely to affect scoring. NG (2009) did not use true but system mentions, and LUO i ZITOUNI (2005) had access to the entire ACE-2004 formal test sets, while the remaining systems, due to licensing restrictions, were evaluated on only a portion of the ACE-2004 training set.

5.3.2 Desiderata for a coreference evaluation measure

Coreference is a type of clustering task, but it is special in that each item in a cluster bears the same relationship, referential identity, with all other items in the same cluster, plus the fact that a large number of clusters are singletons. Thus, only two of the four formal constraints for clustering evaluation metrics pointed out by AMIGÓ *et al.* (2009) apply to coreference. AMIGÓ *et al.* (2009) formal constraints include: (1) cluster homogeneity, i.e., clusters should not mix items belonging to different categories; (2) cluster completeness, i.e., items belonging to the same category should be grouped in the same cluster; (3) rag bag, i.e., it is preferable

GOLD = { {Barack Obama, the president, Obama}, {Sarkozy}, {Berlin}, {the UN}, {today} }

S1 = { {Barack Obama, the president, Obama, Sarkozy}, {Berlin}, {the UN}, {today} }

S2 = { {Barack Obama, the president, Obama}, {Sarkozy, Berlin, the UN, today} }

Figure 5.4: An example not satisfying constraint (3): The output S2 with a rag-bag cluster is equally preferable to S1.

GOLD = { {Barack Obama, the president, Obama}, {the French capital, Paris}, {the Democrats, the Democrats} }

S1 = { {Barack Obama, the president, Obama}, {the French capital}, {Paris}, {the Democrats}, {the Democrats} }

S2 = { {Barack Obama, the president}, {Obama}, {the French capital, Paris}, {the Democrats, the Democrats} }

Figure 5.5: An example not satisfying constraint (4): The output S2 with a small error in a large cluster is equally preferable to S1.

to have clean clusters plus a cluster with miscellaneous items over having clusters with a dominant category plus additional noise; and (4) cluster size versus quantity, i.e., a small error in a large cluster is preferable to a large number of small errors in small clusters.

While the first two constraints undoubtedly hold for coreference resolution, the last two do not necessarily. What makes coreference resolution special with respect to other clustering tasks is the propagation of relations within an entity caused by the transitive property of coreference. That is to say, unlike regular clustering, where assigning a new item to a cluster is a mere question of classifying that item into a specific category, in coreference resolution assigning a new mention to an entity implies that the mention is coreferent with *all* other mentions that have been assigned to that same entity. Thus, the larger an entity is, the more coreferent links will be asserted for each new mention that is added.

To illustrate: to us, given the GOLD in Fig. 5.4, the output produced by system S2 is not better than that produced by system S1, as it would follow from constraint (3). In fact, if the rag-bag entity contained more singletons, including an additional wrong singleton would make S2 even worse than S1. Similarly, in Fig. 5.5, S2 is not better than S1, as constraint (4) suggests.

AMIGÓ *et al.* (2009) show that whereas B^3 satisfies all four constraints, measures based on counting pairs, such as the Rand index, satisfy only constraints (1) and (2). This is a reason why Rand is a good starting point for developing the BLANC measure for coreference resolution in Section 5.4. As described in Section 5.3.1, the three most important points that remain unsolved by the current coreference metrics are:

1. *Singletons*. Since including a mention in the wrong chain hurts P, a correct

decision to NOT link a mention should be rewarded as well. Rewarding correctly identified singletons, however, needs to be moderate, leaving enough margin for the analysis of correctly identified multi-mention entities.

2. *Boundary cases*. Special attention needs to be paid to the behavior of the evaluation measure when a system outputs (1) all singletons, or (2) one entity (i.e., all mentions are linked).

3. *Number of mentions*. The longer the entity chain, the more coreferent mentions it contains, each mention inheriting the information predicated of the other mentions. Thus, a correct large entity should be rewarded more than a correct small entity, and a wrong large entity should be penalized more than a wrong small entity.

We suggest that a good coreference evaluation measure should conform to the following desiderata:

1. Range from 0 for poor performance to 1 for perfect performance.

2. Be monotonic: Solutions that are obviously better should obtain higher scores.

3. Reward P more than R: Stating that two mentions are coreferent when they are not is more harmful than missing a correct coreference link.[5] Hence, the score should move closer to 1 as

   - More correct coreference links are found,
   - more correct singletons are found,
   - fewer wrong coreference links are made.

4. Provide sufficiently fine scoring granularity to allow detailed discrimination between systems across the whole range [0, 1].

5. As nearly as possible, maintain the same degree of scoring granularity throughout the whole range [0, 1].

5.4 BLANC: BiLateral Assessment of Noun-phrase Coreference

In order to facilitate future research, we propose BLANC, a measure obtained by applying the Rand index (RAND, 1971) to coreference and taking into account the above-mentioned problems and desiderata. The class-based methods suffer from the essential problem that they reward each link to a class equally no matter how

[5] Although this is debatable, as it might depend on the application for which the coreference output is used, it is a widespread belief among researchers that P matters more than R in coreference resolution.

large the class is; assigning a mention to a small class is scored equally as assigning it to a large one. But in principle, assigning it to a large one is making a larger number of pairwise decisions, each of which is equally important. Also, singletons well identified are rewarded like correct full multi-mention entities. In addition, the MUC metric suffers from the essential problem that it does not explicitly reward correctly identified singletons, yet penalizes singletons when incorrectly included as part of a chain, while it is too lenient with penalizing wrong coreference links.

5.4.1 Implementing the Rand index for coreference evaluation

From what has been said in Section 5.3, the Rand index seems to be especially adequate for evaluating coreference since it allows us to measure 'non-coreference' as well as coreference links. This makes it possible to correctly handle singletons as well as to reward correct coreference chains commensurately with their length.[6] The interesting property of implementing Rand for coreference is that the sum of all coreference and non-coreference links together is constant for a given set of N mentions, namely the triangular number $N(N-1)/2$. By interpreting a system's output as linking each mention to all other mentions as either coreferent or non-coreferent, we can observe the relative distributions within this constant total of coreference and non-coreference links against the gold standard.

The Rand index (5.1) uses N_{00} (i.e., the number of mention pairs that are in the same entity in both GOLD and SYS) and N_{11} (i.e., the number of mention pairs that are in different entities in both GOLD and SYS) as agreement indicators between the two partitions GOLD and SYS. The value of Rand lies between 0 and 1, with 0 indicating that the two partitions do not agree on any pair of mentions and 1 indicating that the partitions are identical.

$$\text{Rand} = \frac{N_{00} + N_{11}}{N(N-1)/2} \tag{5.1}$$

BLANC borrows the 'bilateral' nature of Rand to take into consideration both coreference links (N_{00}) and non-coreference links (N_{11}), but modifies it such that every decision of coreferentiality is assigned equal importance. Thus, BLANC models coreference resolution better by addressing the significant imbalance between the number of coreferent mentions and singletons observed in real data. Further, whereas class-based metrics need to address the fact that GOLD and SYS might not contain the same number of entities, and the MUC metric focuses on comparing a possibly unequal number of coreference links, BLANC is grounded in the fact that the total number of links remains constant across GOLD and SYS.

5.4.1.1 Coreference and non-coreference links

BLANC is best explained considering two kinds of decisions:

[6] We define a non-coreference link to hold between every two mentions that are deemed to NOT corefer.

		SYS		Sums
		Coreference	Non-coreference	
GOLD	Coreference	*rc*	*wn*	*rc* + *wn*
	Non-coreference	*wc*	*rn*	*wc* + *rn*
Sums		*rc* + *wc*	*wn* + *rn*	*L*

Table 5.4: The BLANC confusion matrix

1. The coreference decisions (made by the coreference system)
   (a) A **coreference link** (*c*) holds between every two mentions that corefer.
   (b) A **non-coreference link** (*n*) holds between every two mentions that do not corefer.
2. The correctness decisions (made by the evaluator)
   (a) A **right link** (*r*) has the same value (coreference or non-coreference) in GOLD and SYS (i.e., when the system is correct).
   (b) A **wrong link** (*w*) does not have the same value (coreference or non-coreference) in GOLD and SYS (i.e., when the system is wrong).

Table 5.4 shows the 2x2 confusion matrix obtained by contrasting the system's coreference decisions against the gold standard decisions. All cells outside the diagonal contain errors of one class being mistaken for the other. BLANC resembles Pairwise F1 as far as coreference links are concerned, but it adds the additional dimension of non-coreference links.

Let N be the total number of mentions in a document d, and let L be the total number of mention pairs (i.e., pairwise links) in d, thereby including both coreference and non-coreference links, then

$$L = N(N-1)/2$$

The total number of links in the SYS partition of d is the sum of the four possible types of links, and it equals L:

$$rc + wc + rn + wn = L$$

where *rc* are the number of right coreference links, *wc* are the number of wrong coreference links, *rn* are the number of right non-coreference links, and *wn* are the number of wrong non-coreference links.

The confusion matrix for the example in Fig. 5.1 is shown in Table 5.5. As the text has fourteen mentions, the total number of links is ninety-one. The system correctly identifies two coreference links (m_5–m_{12}, m_7–m_9), and wrongly another three coreference links (m_4–m_6, m_7–m_{14}, m_9–m_{14}). Every right coreference link that is missed by the system necessarily produces a wrong non-coreference link

		SYS		Sums
		Coreference	Non-coreference	
GOLD	Coreference	2	2	4
	Non-coreference	3	84	87
Sums		5	86	91

Table 5.5: The BLANC confusion matrix for the example in Fig. 5.1

Score	Coreference	Non-coreference	
P	$P_c = \frac{rc}{rc+wc}$	$P_n = \frac{rn}{rn+wn}$	BLANC-P = $\frac{P_c+P_n}{2}$
R	$R_c = \frac{rc}{rc+wn}$	$R_n = \frac{rn}{rn+wc}$	BLANC-R = $\frac{R_c+R_n}{2}$
F	$F_c = \frac{2P_cR_c}{P_c+R_c}$	$F_n = \frac{2P_nR_n}{P_n+R_n}$	BLANC = $\frac{F_c+F_n}{2}$

Table 5.6: Definition: Formula for BLANC

(m_5–m_{14}, m_{12}–m_{14}). The rest are eighty-four right non-coreference links. The confusion matrix shows the balance between coreference and non-coreference links with respect to the gold partition.

The singleton problem pointed out in Section 5.3 becomes evident in Table 5.5: the number of non-coreference links is much higher than the number of coreference links. The class imbalance problem of coreference resolution causes that if the Rand index is applied as originally defined by RAND (1971), it concentrates in a small interval near 1 with hardly any discriminatory power. A chance-corrected Rand index has been proposed (HUBERT i ARABIE, 1985), but it is of no use for the coreference problem, given that the computation of expectation only depends on the number of pairs in the same cluster, thus ignoring singletons.

In order to take into account the under-representation of coreference links in the final BLANC score, we compute P, R, and F separately for the two types of link (coreference and non-coreference) and then average them for the final score. The definition of BLANC is shown in Table 5.6. In BLANC, both coreference and non-coreference links contribute to the final score, but neither more than 50%. BLANC-P and BLANC-R correspond to the average of the two P and R scores, respectively. The final BLANC score corresponds to the average of the two F-scores. Applying the Rand index, the novelty of BLANC resides in putting equal emphasis on coreference and non-coreference links. Table 5.7 shows the different measures under discussion for the example in Fig. 5.1.

MUC-F	B^3-F	CEAF	BLANC
57.14	86.76	85.71	70.78

Table 5.7: Performance of the example in Fig. 5.1

5.4.1.2 Boundary cases

In boundary cases (when for example, SYS or GOLD contain only singletons or only a single set), either P_c or P_n and/or either R_c or R_n are undefined, as one or more denominators will be 0. For these cases we define small variations of the general formula for BLANC shown in Table 5.6.

- If SYS contains a single entity, then it only produces coreference links. If GOLD coincides with SYS, BLANC scores equal 100. If GOLD is fully the dual (i.e., it contains only singletons), BLANC scores equal 0. Finally, if GOLD contains links of both types, P_n, R_n, and F_n equal 0.

- If SYS contains only singletons, then it only produces non-coreference links. If GOLD coincides with SYS, BLANC scores equal 100. If GOLD is fully the dual (i.e., it contains a single entity), BLANC scores equal 0. Finally, if GOLD contains links of both types, P_c, R_c, and F_c equal 0.

- If GOLD includes links of both types but SYS contains no right coreference link, then P_c, R_c, and F_c equal 0. Instead, if SYS contains no right non-coreference link, then P_n, R_n, and F_n equal 0.

- If SYS contains links of both types but GOLD contains a single entity, BLANC scores equal P_c, R_c, and F_c. Instead, if GOLD contains only singletons, BLANC scores equal P_n, R_n, and F_n.

A near-boundary case reveals the main weakness of BLANC. This is the case in which all links but one are non-coreferent and the system outputs only non-coreference links. Then, the fact that BLANC places equal importance on the one link as on all the remaining links together leads to a too severe penalization, as the BLANC score will never be higher than 50. One can either simply accept this as a quirk of BLANC or, following the beta parameter used in the F-score, can introduce a parameter that enables the user to change the relative weights given to coreference and non-coreference links. We provide details in the following section.

5.4.1.3 The α parameter

After analyzing several coreferentially annotated corpora, we found that the average text contains between 60% and 80% singletons (depending on the coding scheme). Thus, simply averaging the coreference and non-coreference scores seems to be the best decision. However, given extraordinary cases like the one presented

at the end of Section 5.4.1.2 or for those researchers that consider it to be convenient, we present the weighted version of BLANC:

$$\text{BLANC}_\alpha = \alpha \text{F}_c + (1-\alpha)\text{F}_n$$

BLANC_α lets users choose the weights they want to put on coreference and non-coreference links. In the default version of BLANC (Table 5.6), α=0.5. Setting α closer to 1 will give a larger weight to coreference links, while setting α closer to 0 will have the opposite effect. For the problematic near-boundary case in which all links but one are non-coreferent in GOLD, evaluating with $\text{BLANC}_{\alpha=0.1}$ will be much less severe than evaluating with the default BLANC.

5.4.2 Identification of mentions

An additional drawback that has been pointed out for class-based metrics like B^3 and CEAF is their assumption of working with true mentions, ignoring the problem of evaluating end-to-end systems, where some mentions in SYS might not be correct; i.e., might not be mapped onto any mention in GOLD and *vice versa*. These are called 'twinless' mentions by STOYANOV *et al.* (2009). BENGTSON i ROTH (2008) simply discard twinless mentions, and RAHMAN i NG (2009) limit to removing only those twinless system mentions that are singletons, as in these cases no penalty should be applied. Recently, CAI i STRUBE (2010) have proposed two variants of B^3 and CEAF that put twinless gold mentions into SYS as singletons and discard singleton twinless system mentions. To calculate P, wrongly resolved twinless system mentions are put into GOLD; to calculate R, only the gold entities are considered.

We agree that proper evaluation of a coreference system should take into account true versus system mentions. However, the mention identification task strictly belongs to syntax as it is closely related to the problem of identifying noun-phrase boundaries, followed by a filtering step in which only referential noun phrases are retained. It is clearly distinct from coreference resolution, whose goal is to link those noun phrases that refer to the same entity. One single metric giving the overall result for the two tasks together is obscure in that it is not informative as to whether a system is very good at identifying coreference links but poor at identifying mention boundaries, or *vice versa*. Therefore, instead of merging the two tasks, we propose to consider mention identification as its own task and separate its evaluation from that of coreference resolution (POPESCU-BELIS *et al.*, 2004). In brief, a measure for each problem is as follows:

- *Mention identification*. This evaluation computes the correctness of the mentions that are being resolved, regardless of the structure of coreference links. Standard P and R are computed to compare the sets of mentions of GOLD and SYS. P is defined as the number of common mentions between GOLD and SYS divided by the number of system mentions; R is defined as the number of common mentions between GOLD and SYS divided by the number of true mentions. Two versions for the matching module are possible:

  - Strict matching. A system mention is considered to be correctly identified when it exactly matches the corresponding gold mention.
  - Lenient matching. A system mention is considered to be correctly identified when it matches at least the head of the corresponding gold mention (and does not include any tokens outside the gold mention).[7]

- *Correctness of coreference*. This evaluation computes the correctness of the coreference links predicted between the mentions shared by GOLD and SYS. The BLANC measure is applied to this set of correctly recognized mentions.

In this way, it might be possible to improve under-performing systems by combining, for instance, the strengths of a system that obtains a high coreference score but a low mention-identification score with the strengths of a system that performs badly in coreference resolution but successfully in the identification of mentions. Similarly, one should not be led to believe that improving the set of coreference features will necessarily result in higher scores, as the system's mention-identification score might reveal that the underlying problem is a poor detection of true mentions.

5.5 Discriminative power

This section empirically demonstrates the power of BLANC by comparing its scores with those of MUC, B^3, CEAF, and the Rand index on both artificial and real gold/system partitions. The insight provided by BLANC is free of the problems noted in Section 5.3. This being said, we need to draw attention to the difficulty of agreeing on what 'correctness' means in coreference resolution. People's intuitions about the extreme boundary cases largely coincide, but those about intermediate cases, which are harder to evaluate, might differ considerably due to the complex trade-off between P and R. Thus, the discussion that follows is based on what we believe to be the best ranking of system responses according to our intuitions and to our experience in coreference annotation and resolution.

5.5.1 Results on artificial data

We take the gold partition in the first row of Table 5.8 as a working example. It is representative of a real case: it contains seventy mentions, 95% singleton entities, a two-mention entity, a three-mention entity, and a four-mention entity. Each number represents a different mention; parentheses identify entities (i.e., they group mentions that corefer); and multi-mention entities are highlighted in bold. Table 5.8 also contains eight sample responses—output by different hypothetical coreference resolution systems—that contain different types of errors. See the decomposition into BLANC's four types of links in Table 5.9, a quantitative representation of the

[7] Lenient matching is equivalent to the MIN attribute used in the MUC guidelines (HIRSCHMAN i CHINCHOR, 1997) to indicate the minimum string that the system under evaluation must include.

Response	Output
$Gold_1$	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62,63,64,65) (66,67,68) (69,70)**
System A	**(1,2)** (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62,63,64,65) (66,67,68) (69,70)**
System B	**(1,62,63,64,65)** (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(66,67,68) (69,70)**
System C	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62,63,64,65)** (66) (67) (68) **(69,70)**
System D	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62,63,64,65,66,67,68) (69,70)**
System E	**(1,62,63)** (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) **(28,64,65)** (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(66,67,68) (69,70)**
System F	**(1,62)** (2) (3) **(4,63)** (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) **(28,64)** (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) **(57,65)** (58) (59) (60) (61) **(66,67,68) (69,70)**
System G	All singletons
System H	One entity

Table 5.8: Different system responses for a gold standard $Gold_1$

System	#entities	#singletons	*rc*	*rn*	*wc*	*wn*
A	63	59	10	2,404	1	0
B	63	60	10	2,401	4	0
C	66	64	7	2,405	0	3
D	63	61	10	2,393	12	0
E	63	59	6	2,401	4	4
F	63	57	4	2,401	4	6
G	70	70	0	2,405	0	10
H	1	0	10	0	2,405	0

Table 5.9: Decomposition of the system responses in Table 5.8

System	MUC-F	B^3-F	CEAF	RAND	BLANC
A	92.31	99.28	98.57	99.96	97.61
B	92.31	98.84	98.57	99.83	91.63
C	80.00	98.55	97.14	99.88	91.15
D	92.31	97.49	95.71	99.50	81.12
E	76.92	96.66	95.71	99.67	79.92
F	46.15	94.99	94.29	99.59	72.12
G	—	95.52	91.43	99.59	49.90
H	16.00	3.61	5.71	0.41	0.41

Table 5.10: Performance of the systems in Table 5.8

quality of the systems given in Table 5.8. The responses are ranked in order of quality, from the most accurate response to the least (response A is better than response B, B is better than C, and so on, according to our intuitions[8]).

System A commits only one P error by linking two non-coreferent mentions; system B looks similar to A but is worse in that a singleton is clustered in a four-mention entity, thus producing not one but four P errors. System C exhibits no P errors but is weak in terms of R, as it fails to identify a three-mention entity. Although system D is clean in terms of R, it suffers from a severe P problem due to the fusion of the three- and four-mention entities in one large entity. System E is worse than the previous responses in that it shows both P and R errors: the four-mention entity is split into two and a singleton is added to both of them. System F worsens the previous output by completely failing to identify the four-mention entity and creating four incorrect two-mention entities. Finally, systems G and H represent the two boundary cases, the former being preferable to the latter, since at least it gets the large number of singletons, while the latter has a serious problem in P.

The performance of these system responses according to different measures is

[8]Readers and reviewers of this section frequently comment that this ranking is not clearly apparent; other variations seem equally good. We concede this readily. We argue that in cases when several rankings seem intuitively equivalent to people, one can accept the ranking of a metric, as long as it assigns relatively close scores to the equivalent cases.

System	MUC		B^3		CEAF	BLANC	
	P	R	P	R	P/R	P	R
A	85.71	100.00	98.57	100.00	98.57	95.45	99.98
B	85.71	100.00	97.71	100.00	98.57	85.71	99.92
C	100.00	66.67	100.00	97.14	97.14	99.94	85.00
D	85.71	100.00	95.10	100.00	95.71	72.73	99.75
E	71.43	83.33	96.19	97.14	95.71	79.92	79.92
F	42.86	50.00	94.29	95.71	94.29	74.88	69.92
G	—	—	100.00	91.43	91.43	49.79	50.00
H	8.70	100.00	1.84	100.00	5.71	0.21	50.00

Table 5.11: P and R scores for the systems in Table 5.8

given in Tables 5.10 and 5.11. In them, we can see how BLANC addresses the three problems noted in Section 5.3.2.

1. *Singletons*. The BLANC score decreases as the response quality decreases. It successfully captures the desired ranking, so does CEAF (although with fewer distinctions, see the 'number of mentions' problem below), and so does B^3 if we leave aside the boundary responses G and H. BLANC, however, shows a much wider interval (from 97.61% to 49.90%) than CEAF (from 98.57% to 91.43%) and B^3 (from 99.28% to 94.99%), thus providing a larger margin of variation, and a finer granularity. The singleton problem is solved by rewarding the total number of correct singletons as much as the total number of correct mentions in multi-mention entities. Note that the original Rand index makes it impossible to discriminate between systems and it does not even rank them as intuitively expected.

2. *Boundary cases*. MUC fails to capture the fact that the all-singletons response G is better than the one-entity response H. On the other hand, B^3 and CEAF give a score close to 0% for H, yet close to 100% for G. It is counterintuitive that a *coreference* resolution system that outputs as many entities as mentions—meaning that it is doing nothing—gets such a high score. BLANC successfully handles the boundary responses by setting an upper bound of 50% on R.

3. *Number of mentions*. The fact that MUC and CEAF give the same score to responses A and B shows their failure at distinguishing that the latter is more harmful than the former, as it creates more false coreference links. Namely, the information predicated of mention 1 is extended to mentions 61, 62, 63, and 64, and reciprocally mention 1 gets all the information predicated of mentions 61, 62, 63, and 64. Similarly, CEAF does not distinguish response D from E. In contrast, BLANC can discriminate between these responses because its reward of multi-mention entities is correlated with the number of coreference links contained in them.

Response	Output
Gold_2	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) **(17,18)**
System A	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18)
System B	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) **(16,17)** (18)
System C	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) **(15,16)** (17) (18)
System D	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) **(15,16) (17,18)**

Table 5.12: Different system responses for a gold standard Gold_2

System	MUC-F	B^3-F	CEAF	$\text{BLANC}_{\alpha=0.5}$	$\text{BLANC}_{\alpha=0.2}$	$\text{BLANC}_{\alpha=0.1}$
A	—	97.14	94.44	49.84	79.74	89.70
B	0.00	94.44	94.44	49.67	79.47	89.41
C	0.00	94.44	88.89	49.67	79.47	89.41
D	66.67	97.14	94.44	83.17	93.07	96.37

Table 5.13: Performance for the systems in Table 5.12

The constructed example in Table 5.12 serves to illustrate BLANC's major weakness, which we discussed at the end of Section 5.4.1.2. Performance is presented in Table 5.13. Notice the enormous leap between the $\text{BLANC}_{\alpha=0.5}$ score for system D and the other three. This is due to the fact that partitions A, B, and C contain no right coreference link, and so BLANC is equal to the correctness of non-coreference links divided by two. The α parameter introduced in Section 5.4.1.3 is especially adequate for this type of cases. The difference in the scores for D and the rest of systems diminishes when α=0.2 or α=0.1 (the two last columns).

This same example, in fact, reveals weaknesses of all the measures. Owing to the fact that the MUC score does not reward correctly identified singletons, it is not able to score the first three responses, thus showing even a larger rise in response D. The B^3 and CEAF measures score responses A and D the same, but only the latter succeeds in identifying the only coreference link that exists in the truth—a very relevant fact given that the ultimate goal of a coreference resolution system is not outputting only singletons (as system A does) but solving coreference. Finally, it is puzzling that CEAF considers response B to be appreciably better than response C—they are scored the same by B^3 and BLANC. This is a weakness due to CEAF's one-to-one alignment: In B, the three final entities find a counterpart in the gold standard, whereas in C, only one of the two final entities gets mapped.

5.5.2 Results on real data

In order not to reach conclusions solely derived from constructed toy examples, we run a prototype learning-based coreference resolution system—inspired by SOON *et al.* (2001), NG i CARDIE (2002*b*), and LUO *et al.* (2004)—on 33 documents of the ACE-2004 corpus. A total of five different resolution models are tried to

Resolution model	MUC-F	B^3-F	CEAF	BLANC
A. All-singletons baseline	—	67.51	50.96	48.61
B. Head-match baseline	52.93	76.60	66.46	66.35
C. Strong match	64.69	75.56	**70.63**	**73.76**
D. Best match	61.60	**76.76**	69.19	71.98
E. Weak match	**70.34**	70.24	64.00	66.50

Table 5.14: Different coreference resolution models run on ACE-2004

Resolution model	#entities	#singletons	*rc*	*rn*	*wc*	*wn*
A. All-singletons baseline	1,464	1,464	0	39,672	0	2,272
B. Head-match baseline	1,124	921	506	39,560	112	1,766
C. Strong match	735	400	1,058	38,783	889	1,214
D. Best match	867	577	870	39,069	603	1,402
E. Weak match	550	347	1,757	34,919	4,753	515

Table 5.15: Decomposition of the system responses in Table 5.14

enable a richer analysis and comparison between the different evaluation metrics. The results are presented in Table 5.14. For a detailed analysis we address the reader to RECASENS i HOVY (2010).

The first two are baselines that involve no learning: model A is the all-singletons baseline, and B clusters in the same entity all the mentions that share the same head. In C, D, and E, a pairwise coreference classifier is learnt (i.e., given two mentions, it classifies them as either coreferent or non-coreferent). In C and D, whenever the classifier considers two mentions to be coreferent and one of them has already been clustered in a multi-mention entity, the new mention is only clustered in that same entity if all pairwise classifications with the other mentions of the entity are also classified as coreferent. The difference between C and D lies in the initial mention pairs that form the basis for the subsequent process: C takes the first mention in textual order that is classified as coreferent with the mention under consideration, while D takes the mention that shows the highest confidence among the previous. E is a simplified version of C that performs no additional pairwise checks.

The best way to judge the quality of each response is to look at the actual data, but space limitations make this impossible. However, we can gain an approximation by looking at Table 5.15, which shows the number of entities output by each system and how many are singletons, as well as the number of correct and incorrect links of each type. Note that high numbers in the *wc* column indicate poor P, whereas high numbers in the *wn* column indicate poor R. Although the trade-off between P and R makes it hard to reach a conclusion as to whether C or D should be ranked first, the low quality of A, and especially E, is an easier conclusion to reach. The head-match baseline achieves high P but low R.

If we go back to Table 5.14, we can see that no two measures produce the same ranking of systems. The severe problems behind the MUC score are again

System	MUC	B^3	CEAF	ACE-value	BLANC
			ACE-2		
All-singletons baseline	—	55.9	38.8		**47.8**
One-entity baseline	76.5	17.3	21.7		**7.8**
LUO *et al.* (2004)	80.7	77.0	73.2	89.8	**77.2**
			ACE-2004		
All-singletons baseline	—	59.0	41.8		**48.1**
One-entity baseline	74.4	17.8	21.4		**7.0**
LUO i ZITOUNI (2005)	86.0	83.7	82.0	91.6	**81.4**
BENGTSON i ROTH (2008)	75.8	80.8	75.0		**75.6**

Table 5.16: Performance of state-of-the-art systems on ACE according to BLANC

manifested: it ranks model E first because it identifies a high number of coreference links, despite containing many incorrect ones. This model produces an output that is not satisfactory because it tends to overmerge. The fact that B^3 ranks D and B first indicates its focus on P rather than R. Thus, B^3 tends to score best those models that are more conservative and that output a large number of singletons. Finally, CEAF and BLANC agree in ranking C the best. An analysis of the data also supports the idea that strong match achieves the best trade-off between P and R.

Similar problems with the currently used evaluation metrics were also shown by the six systems that participated in the SemEval-2010 Task 1 on 'Coreference Resolution in Multiple Languages' (RECASENS *et al.*, 2010*b*), where the BLANC measure was publicly used for the first time. Unlike ACE, mentions were not restricted to any semantic type, and the B^3 and CEAF scores for the all-singletons baseline were hard to beat even by the highest performing systems. The BLANC scores, in contrast, tended to stay low regardless of the number of singletons in the corpus. However, it was not possible to draw definite conclusions about the SemEval shared task because each measure ranked the participating systems in a different order.

Finally, in Table 5.16 we reproduce Table 5.3 adding the BLANC score for the performance of state-of-the-art systems and the all-singletons and one-entity baselines. We can only include the results for those systems whose output responses were provided to us by the authors. It is worth noting that BLANC is closer to B^3 when using the ACE-2 corpus but closer to CEAF when using the ACE-2004 corpus, which is probably due to the different distribution of singletons and multi-mention entities in each corpus. Knowing the state of the art in terms of BLANC will enable future researchers on coreference resolution to compare their performance against these results.

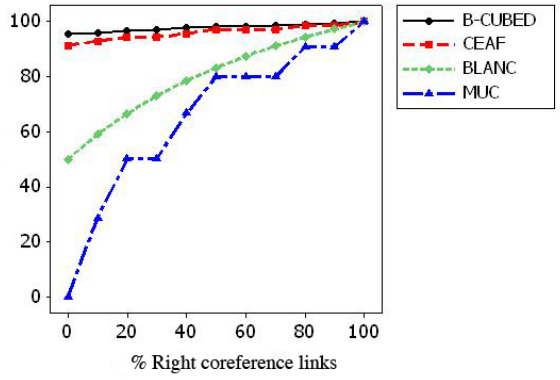

Figure 5.6: The BLANC score curve as the number of right coreference links increases

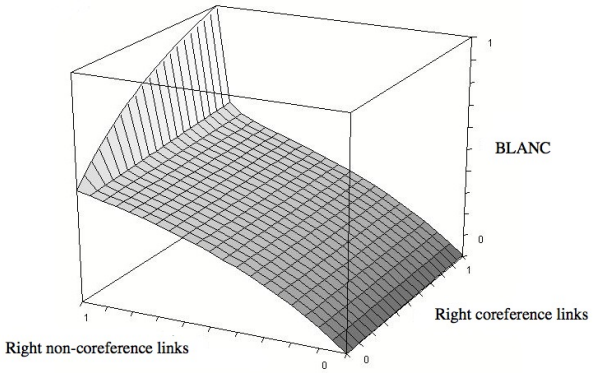

Figure 5.7: The BLANC score surface as a function of right coreference and right non-coreference links, for data from Table 5.8

5.5.3 Plots

A graph plotting the BLANC slope as the percentage of correct coreference links (*rc*) increase is depicted in Fig. 5.6, where the slopes of B^3, CEAF, and MUC are also plotted. The curve slope for BLANC gradually increases, and stays between the other measures, higher than MUC but lower than B^3 and CEAF, which show an almost flat straight line. The 'pinching' of scores close to 100% by B^3 and CEAF is clearly apparent. A coreference resolution system can obtain very high B^3 and CEAF scores (due to the high number of singletons that are present in the gold partition), leaving a too small margin for the evaluation of coreference proper.

We illustrate in Fig. 5.7 the dependency of the BLANC score on degrees of coreference and non-coreference. Fig. 5.7 plots the scores for the example in Table 5.8. The left rear face of the cube—where the right non-coreference (i.e., *rn*) level is a constant 1 and right coreference (*rc*) ranges from zero to 1—displays the

BLANC curve from Fig. 5.6. The front face of the cube shows how—for a constant right coreference of 1—the BLANC score ranges from near zero to 0.5 as right non-coreference ranges from zero to 1. The bend in the surface occurs due to the asymmetry in the number of true coreferences: the smaller the proportion of coreference links to non-coreference links, the sharper the bend and the closer it is to the left face. Systems must achieve correctness of almost all coreference *and* non-coference links to approach the steep curve.

5.6 Conclusion

This article seeks to shed light on the problem of coreference resolution evaluation by providing desiderata for coreference evaluation measures, pointing out the strong and weak points of the main measures that have been used, and proposing the BLANC metric, an implementation of the Rand index for coreference, to provide some further insight on a system's performance. The decomposition into four types of links gives an informative analysis of a system. BLANC fulfills the five desiderata and addresses to some degree the reported shortcomings of the existing measures. Despite its shortcomings, discussed in Sections 5.4.1.2 and 5.5.1, it overcomes the problem of singletons, which we illustrate here for the first time.

The simplicity of the BLANC measure derives from the fact that the sum of the coreference and non-coreference links in the gold and system partitions is the same. Unlike the Rand index, BLANC is the average of two F-scores, one for the coreference links and the other for the non-coreference links. Being two harmonic means, each F-score is lower than the normal average of P and R—unless both are high. As a result, a coreference resolution system has to get *both* P and R for both coreference and non-coreference correct simultaneously to score well under BLANC. Although coreference and non-coreference are duals, ignoring one of the two halves means that some portion of the full link set remains unconsidered by the existing measures.

Tests on artificial and real data show that no evaluation measure is free of weaknesses and so at least two scoring measures should be used when evaluating a system. We argue that BLANC is consistent and achieves a good compromise between P and R. Its discriminative power—higher with respect to currently used metrics like MUC and B^3—facilitates comparisons between coreference resolution systems.

Finally, this article illustrates the need for a fuller comparison of all the evaluation measures, considering corrections required for chance variation, typical variances of scores under different conditions and data sizes, etc. Such a study has not yet been done for any of the measures, and could make a major contribution to the growing understanding of evaluation in the various branches of natural language engineering in general.

Acknowledgments We would like to thank the anonymous reviewers for their helpful questions and comments. We are also indebted to Aron Culotta, Hal Daumé III, Jenny Finkel, Aria Haghighi, Xiaoqiang Luo, Andrew McCallum, Vincent Ng, Hoifung Poon, and Nicholas Rizzolo for answering our request to recompute the performance of their coreference resolution systems with other metrics and/or providing us their system responses. Many thanks to Edgar Gonzàlez for implementation solutions.

This research was partially supported by the Spanish Ministry of Education through an FPU scholarship (AP2006-00994) and the TEXT-MESS 2.0 Project (TIN2009-13391-C04-04).

CAPÍTOL 6

SemEval-2010 Task 1: Coreference Resolution in Multiple Languages

Marta Recasens[⋆], Lluís Màrquez[⋆⋆], Emili Sapena[⋆⋆], M. Antònia Martí[⋆], Mariona Taulé[⋆], Véronique Hoste[†], Massimo Poesio[◇], and Yannick Versley[‡]

[⋆]University of Barcelona
[⋆⋆] Technical University of Catalonia
[†] University College Ghent
[◇] University of Essex/University of Trento
[‡] University of Tübingen

Published in *Proceedings of the ACL 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 1–8, Uppsala, Sweden

Abstract This paper presents the SemEval-2010 task on *Coreference Resolution in Multiple Languages*. The goal was to evaluate and compare automatic coreference resolution systems for six different languages (Catalan, Dutch, English, German, Italian, and Spanish) in four evaluation settings and using four different metrics. Such a rich scenario had the potential to provide insight into key issues concerning coreference resolution: (i) the portability of systems across languages, (ii) the relevance of different levels of linguistic information, and (iii) the behavior of scoring metrics.

6.1 Introduction

The task of coreference resolution, defined as the identification of the expressions in a text that refer to the same discourse entity (1), has attracted considerable attention within the NLP community.

(1) *Major League Baseball* sent *its* head of security to Chicago to review the second incident of an on-field fan attack in the last seven months. *The league* is reviewing security at all ballparks to crack down on spectator violence.

Using coreference information has been shown to be beneficial in a number of NLP applications including Information Extraction (MCCARTHY i LEHNERT, 1995), Text Summarization (STEINBERGER *et al.*, 2007), Question Answering (MORTON, 1999), and Machine Translation. There have been a few evaluation campaigns on coreference resolution in the past, namely MUC (HIRSCHMAN i CHINCHOR, 1997), ACE (DODDINGTON *et al.*, 2004), and ARE (ORASAN *et al.*, 2008), yet many questions remain open:

- To what extent is it possible to implement a general coreference resolution system portable to different languages? How much language-specific tuning is necessary?

- How helpful are morphology, syntax and semantics for solving coreference relations? How much preprocessing is needed? Does its quality (perfect linguistic input versus noisy automatic input) really matter?

- How (dis)similar are different coreference evaluation metrics—MUC, B^3, CEAF and BLANC? Do they all provide the same ranking? Are they correlated?

Our goal was to address these questions in a shared task. Given six datasets in Catalan, Dutch, English, German, Italian, and Spanish, the task we present involved automatically detecting full coreference chains—composed of named entities (NEs), pronouns, and full noun phrases—in four different scenarios. For more information, the reader is referred to the task website.[1]

The rest of the paper is organized as follows. Section 6.2 presents the corpora from which the task datasets were extracted, and the automatic tools used to preprocess them. In Section 6.3, we describe the task by providing information about the data format, evaluation settings, and evaluation metrics. Participating systems are described in Section 6.4, and their results are analyzed and compared in Section 6.5. Finally, Section 6.6 concludes.

6.2 Linguistic resources

In this section, we first present the sources of the data used in the task. We then describe the automatic tools that predicted input annotations for the coreference resolution systems.

[1] http://stel.ub.edu/semeval2010-coref

6.2.1 Source corpora

Catalan and Spanish The AnCora corpora (RECASENS i MARTÍ, 2010) consist of a Catalan and a Spanish treebank of 500k words each, mainly from newspapers and news agencies (El Periódico, EFE, ACN). Manual annotation exists for arguments and thematic roles, predicate semantic classes, NEs, WordNet nominal senses, and coreference relations. AnCora are freely available for research purposes.

Dutch The KNACK-2002 corpus (HOSTE i DE PAUW, 2006) contains 267 documents from the Flemish weekly magazine Knack. They were manually annotated with coreference information on top of semi-automatically annotated PoS tags, phrase chunks, and NEs.

English The OntoNotes Release 2.0 corpus (PRADHAN *et al.*, 2007*a*) covers newswire and broadcast news data: 300k words from The Wall Street Journal, and 200k words from the TDT-4 collection, respectively. OntoNotes builds on the Penn Treebank for syntactic annotation and on the Penn PropBank for predicate argument structures. Semantic annotations include NEs, words senses (linked to an ontology), and coreference information. The OntoNotes corpus is distributed by the Linguistic Data Consortium.[2]

German The TüBa-D/Z corpus (HINRICHS *et al.*, 2005) is a newspaper treebank based on data taken from the daily issues of "die tageszeitung" (taz). It currently comprises 794k words manually annotated with semantic and coreference information. Due to licensing restrictions of the original texts, a taz-DVD must be purchased to obtain a license.[2]

Italian The LiveMemories corpus (RODRÍGUEZ *et al.*, 2010) will include texts from the Italian Wikipedia, blogs, news articles, and dialogues (MapTask). They are being annotated according to the ARRAU annotation scheme with coreference, agreement, and NE information on top of automatically parsed data. The task dataset included Wikipedia texts already annotated.

The datasets that were used in the task were extracted from the above-mentioned corpora. Table 6.1 summarizes the number of documents (docs), sentences (sents), and tokens in the training, development and test sets.[3]

[2]Free user license agreements for the English and German task datasets were issued to the task participants.

[3]The German and Dutch training datasets were not completely stable during the competition period due to a few errors. Revised versions were released on March 2 and 20, respectively. As to the test datasets, the Dutch and Italian documents with formatting errors were corrected after the evaluation period, with no variations in the ranking order of systems.

	Training			Development			Test		
	#docs	#sents	#tokens	#docs	#sents	#tokens	#docs	#sents	#tokens
Catalan	829	8,709	253,513	142	1,445	42,072	167	1,698	49,260
Dutch	145	2,544	46,894	23	496	9,165	72	2,410	48,007
English	229	3,648	79,060	39	741	17,044	85	1,141	24,206
German	900	19,233	331,614	199	4,129	73,145	136	2,736	50,287
Italian	80	2,951	81,400	17	551	16,904	46	1,494	41,586
Spanish	875	9,022	284,179	140	1,419	44,460	168	1,705	51,040

Table 6.1: Size of the task datasets

6.2.2 Preprocessing systems

Catalan, Spanish, English Predicted lemmas and PoS were generated using FreeLing[4] for Catalan/Spanish and SVMTagger[5] for English. Dependency information and predicate semantic roles were generated with JointParser, a syntactic-semantic parser.[6]

Dutch Lemmas, PoS and NEs were automatically provided by the memory-based shallow parser for Dutch (DAELEMANS *et al.*, 1999), and dependency information by the Alpino parser (VAN NOORD *et al.*, 2006).

German Lemmas were predicted by TreeTagger (SCHMID, 1995), PoS and morphology by RFTagger (SCHMID i LAWS, 2008), and dependency information by MaltParser (HALL i NIVRE, 2008).

Italian Lemmas and PoS were provided by TextPro,[7] and dependency information by MaltParser.[8]

6.3 Task description

Participants were asked to develop an automatic system capable of assigning a discourse entity to every mention,[9] thus identifying all the NP mentions of every discourse entity. As there is no standard annotation scheme for coreference and the source corpora differed in certain aspects, the coreference information of the task datasets was produced according to three criteria:

- Only NP constituents and possessive determiners can be mentions.

[4] http://www.lsi.upc.es/ nlp/freeling

[5] http://www.lsi.upc.edu/ nlp/SVMTool

[6] http://www.lsi.upc.edu// xlluis/?x=cat:5

[7] http://textpro.fbk.eu

[8] http://maltparser.org

[9] Following the terminology of the ACE program, a *mention* is defined as an instance of reference to an object, and an *entity* is the collection of mentions referring to the same object in a document.

- Mentions must be referential expressions, thus ruling out nominal predicates, appositives, expletive NPs, attributive NPs, NPs within idioms, etc.
- Singletons are also considered as entities (i.e., entities with a single mention).

To help participants build their systems, the task datasets also contained both gold-standard and automatically predicted linguistic annotations at the morphological, syntactic and semantic levels. Considerable effort was devoted to provide participants with a common and relatively simple data representation for the six languages.

6.3.1 Data format

The task datasets as well as the participants' answers were displayed in a uniform column-based format, similar to the style used in previous CoNLL shared tasks on syntactic and semantic dependencies (2008/2009).[10] Each dataset was provided as a single file per language. Since coreference is a linguistic relation at the discourse level, documents constitute the basic unit, and are delimited by "#begin document ID" and "#end document ID" comment lines. Within a document, the information of each sentence is organized vertically with one token per line, and a blank line after the last token of each sentence. The information associated with each token is described in several columns (separated by "\t" characters) representing the following layers of linguistic annotation.

ID (column 1). Token identifiers in the sentence.

Token (column 2). Word forms.

Lemma (column 3). Token lemmas.

PoS (column 5). Coarse PoS.

Feat (column 7). Morphological features (PoS type, number, gender, case, tense, aspect, etc.) separated by a pipe character.

Head (column 9). ID of the syntactic head ("0" if the token is the tree root).

DepRel (column 11). Dependency relations corresponding to the dependencies described in the Head column ("sentence" if the token is the tree root).

NE (column 13). NE types in open-close notation.

Pred (column 15). Predicate semantic class.

APreds (column 17 and subsequent ones). For each predicate in the Pred column, its semantic roles/dependencies.

Coref (last column). Coreference relations in open-close notation.

The above-mentioned columns are "gold-standard columns," whereas columns 4, 6, 8, 10, 12, 14, 16 and the penultimate contain the same information as the respective previous column but automatically predicted—using the preprocessing systems listed in Section 6.2.2. Neither all layers of linguistic annotation nor all

ID	Token	Intermediate columns	Coref
1	Major	...	(1
2	League	...	_
3	Baseball	...	1)
4	sent	...	_
5	its	...	(1)|(2
6	head	...	_
7	of	...	_
8	security	...	(3)|2)
9	to	...	_
...	...	...	...
27	The	...	(1
28	league	...	1)
29	is	...	_

Table 6.2: Format of the coreference annotations (corresponding to example (1) in Section 6.1)

gold-standard and predicted columns were available for all six languages (underscore characters indicate missing information).

The coreference column follows an open-close notation with an entity number in parentheses (see Table 6.2). Every entity has an ID number, and every mention is marked with the ID of the entity it refers to: an opening parenthesis shows the beginning of the mention (first token), while a closing parenthesis shows the end of the mention (last token). For tokens belonging to more than one mention, a pipe character is used to separate multiple entity IDs. The resulting annotation is a well-formed nested structure (CF language).

6.3.2 Evaluation settings

In order to address our goal of studying the effect of different levels of linguistic information (preprocessing) on solving coreference relations, the test was divided into four evaluation settings that differed along two dimensions.

Gold-standard versus **Regular setting.** Only in the gold-standard setting were participants allowed to use the gold-standard columns, including the last one (of the test dataset) with true mention boundaries. In the regular setting, they were allowed to use only the automatically predicted columns. Obtaining better results in the gold setting would provide evidence for the relevance of using high-quality preprocessing information. Since not all columns were available for all six languages, the gold setting was only possible for Catalan, English, German, and Spanish.

[10] http://www.cnts.ua.ac.be/conll2008

Closed versus **Open setting.** In the closed setting, systems had to be built strictly with the information provided in the task datasets. In contrast, there was no restriction on the resources that participants could utilize in the open setting: systems could be developed using any external tools and resources to predict the preprocessing information, e.g., WordNet, Wikipedia, etc. The only requirement was to use tools that had not been developed with the annotations of the test set. This setting provided an open door into tools or resources that improve performance.

6.3.3 Evaluation metrics

Since there is no agreement at present on a standard measure for coreference resolution evaluation, one of our goals was to compare the rankings produced by four different measures. The task scorer provides results in the two mention-based metrics B^3 (BAGGA i BALDWIN, 1998) and CEAF-ϕ_3 (LUO, 2005), and the two link-based metrics MUC (VILAIN *et al.*, 1995) and BLANC (Recasens and Hovy, to appear). The first three measures have been widely used, while BLANC is a proposal of a new measure interesting to test.

The mention detection subtask is measured with recall, precision, and F_1. Mentions are rewarded with 1 point if their boundaries coincide with those of the gold NP, with 0.5 points if their boundaries are within the gold NP including its head, and with 0 otherwise.

6.4 Participating systems

A total of twenty-two participants registered for the task and downloaded the training materials. From these, sixteen downloaded the test set but only six (out of which two task organizers) submitted valid results (corresponding to eight system runs or variants). These numbers show that the task raised considerable interest but that the final participation rate was comparatively low (slightly below 30%).

The participating systems differed in terms of architecture, machine learning method, etc. Table 6.3 summarizes their main properties. Systems like BART and Corry support several machine learners, but Table 6.3 indicates the one used for the SemEval run. The last column indicates the external resources that were employed in the open setting, thus it is empty for systems that participated only in the closed setting. For more specific details we address the reader to the system description papers in ERK i STRAPPARAVA (2010).

6.5 Results and evaluation

Table 6.4 shows the results obtained by two naive baseline systems: (i) SINGLETONS considers each mention as a separate entity, and (ii) ALL-IN-ONE groups all the mentions in a document into a single entity. These simple baselines reveal limitations of the evaluation metrics, like the high scores of CEAF and B^3 for SIN-

	System Architecture	ML Methods	External Resources
BART (BROSCHEIT *et al.*, 2010)	Closest-first with entity-mention model (English), Closest-first model (German, Italian)	MaxEnt (English, German), Decision trees (Italian)	GermaNet & gazetteers (German), I-Cab gazetteers (Italian), Berkeley parser, Stanford NER, WordNet, Wikipedia name list, U.S. census data (English)
Corry (URYUPINA, 2010)	ILP, Pairwise model	SVM	Stanford parser & NER, WordNet, U.S. census data
RelaxCor (SAPENA *et al.*, 2010)	Graph partitioning (solved by relaxation labeling)	Decision trees, Rules	WordNet
SUCRE (KOBDANI i SCHÜTZE, 2010)	Best-first clustering, Relational database model, Regular feature definition language	Decision trees, Naive Bayes, SVM, MaxEnt	—
TANL-1 (ATTARDI *et al.*, 2010)	Highest entity-mention similarity	MaxEnt	PoS tagger (Italian)
UBIU (ZHEKOVA i KÜBLER, 2010)	Pairwise model	MBL	—

Table 6.3: Main characteristics of the participating systems

GLETONS. Interestingly enough, the naive baseline scores turn out to be hard to beat by the participating systems, as Table 6.5 shows. Similarly, ALL-IN-ONE obtains high scores in terms of MUC. Table 6.4 also reveals differences between the distribution of entities in the datasets. Dutch is clearly the most divergent corpus mainly due to the fact that it only contains singletons for NEs.

Table 6.5 displays the results of all systems for all languages and settings in the four evaluation metrics (the best scores in each setting are highlighted in bold). Results are presented sequentially by language and setting, and participating systems are ordered alphabetically. The participation of systems across languages and settings is rather irregular,[11] thus making it difficult to draw firm conclusions about the aims initially pursued by the task. In the following, we summarize the most relevant outcomes of the evaluation.

Regarding languages, English concentrates the most participants (fifteen entries), followed by German (eight), Catalan and Spanish (seven each), Italian (five), and Dutch (three). The number of languages addressed by each system ranges from one (Corry) to six (UBIU and SUCRE); BART and RelaxCor addressed three languages, and TANL-1 five. The best overall results are obtained for English followed by German, then Catalan, Spanish and Italian, and finally Dutch. Apart from

[11] Only 45 entries in Table 6.5 from 192 potential cases.

	CEAF			MUC			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
SINGLETONS: Each mention forms a separate entity.												
Catalan	61.2	61.2	61.2	0.0	0.0	0.0	61.2	100	75.9	50.0	48.7	49.3
Dutch	34.5	34.5	34.5	0.0	0.0	0.0	34.5	100	51.3	50.0	46.7	48.3
English	71.2	71.2	71.2	0.0	0.0	0.0	71.2	100	83.2	50.0	49.2	49.6
German	75.5	75.5	75.5	0.0	0.0	0.0	75.5	100	86.0	50.0	49.4	49.7
Italian	71.1	71.1	71.1	0.0	0.0	0.0	71.1	100	83.1	50.0	49.2	49.6
Spanish	62.2	62.2	62.2	0.0	0.0	0.0	62.2	100	76.7	50.0	48.8	49.4
ALL-IN-ONE: All mentions are grouped into a single entity.												
Catalan	11.8	11.8	11.8	100	39.3	56.4	100	4.0	7.7	50.0	1.3	2.6
Dutch	19.7	19.7	19.7	100	66.3	79.8	100	8.0	14.9	50.0	3.2	6.2
English	10.5	10.5	10.5	100	29.2	45.2	100	3.5	6.7	50.0	0.8	1.6
German	8.2	8.2	8.2	100	24.8	39.7	100	2.4	4.7	50.0	0.6	1.1
Italian	11.4	11.4	11.4	100	29.0	45.0	100	2.1	4.1	50.0	0.8	1.5
Spanish	11.9	11.9	11.9	100	38.3	55.4	100	3.9	7.6	50.0	1.2	2.4

Table 6.4: Baseline scores

differences between corpora, there are other factors that might explain this ranking: (i) the fact that most of the systems were originally developed for English, and (ii) differences in corpus size (German having the largest corpus, and Dutch the smallest).

Regarding systems, there are no clear "winners." Note that no language-setting was addressed by all six systems. The BART system, for instance, is either on its own or competing against a single system. It emerges from partial comparisons that SUCRE performs the best in *closed*×*regular* for English, German, and Italian, although it never outperforms the CEAF or B^3 singleton baseline. While SUCRE always obtains the best scores according to MUC and BLANC, RelaxCor and TANL-1 usually win based on CEAF and B^3. The Corry system presents three variants optimized for CEAF (Corry-C), MUC (Corry-M), and BLANC (Corry-B). Their results are consistent with the bias introduced in the optimization (see English:*open*×*gold*).

Depending on the evaluation metric then, the rankings of systems vary with considerable score differences. There is a significant positive correlation between CEAF and B^3 (Pearson's $r = 0.91$, $p < 0.01$), and a significant lack of correlation between CEAF and MUC in terms of recall (Pearson's $r = 0.44$, $p < 0.01$). This fact stresses the importance of defining appropriate metrics (or a combination of them) for coreference evaluation.

Finally, regarding evaluation settings, the results in the *gold* setting are significantly better than those in the *regular*. However, this might be a direct effect of the mention recognition task. Mention recognition in the regular setting falls more than 20 F_1 points with respect to the gold setting (where correct mention boundaries were given). As for the *open* versus *closed* setting, there is only one system, RelaxCor for English, that addressed the two. As expected, results show a slight improvement from *closed*×*gold* to *open*×*gold*.

	Mention detection			CEAF			MUC			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
Catalan															
closed×gold															
RelaxCor	100	100	100	70.5	70.5	**70.5**	29.3	77.3	42.5	68.6	95.8	**79.9**	56.0	81.8	59.7
SUCRE	100	100	100	68.7	68.7	68.7	54.1	58.4	**56.2**	76.6	77.4	77.0	72.4	60.2	**63.6**
TANL-1	100	96.8	98.4	66.0	63.9	64.9	17.2	57.7	26.5	64.4	93.3	76.2	52.8	79.8	54.4
UBIU	75.1	96.3	84.4	46.6	59.6	52.3	8.8	17.1	11.7	47.8	76.3	58.8	51.6	57.9	52.2
closed×regular															
SUCRE	75.9	64.5	69.7	51.3	43.6	47.2	44.1	32.3	**37.3**	59.6	44.7	51.1	53.9	55.2	**54.2**
TANL-1	83.3	82.0	82.7	57.5	56.6	**57.1**	15.2	46.9	22.9	55.8	76.6	**64.6**	51.3	76.2	51.0
UBIU	51.4	70.9	59.6	33.2	45.7	38.4	6.5	12.6	8.6	32.4	55.7	40.9	50.2	53.7	47.8
open×gold															
open×regular															
Dutch															
closed×gold															
SUCRE	100	100	100	58.8	58.8	**58.8**	65.7	74.4	**69.8**	65.0	69.2	**67.0**	69.5	62.9	**65.3**
closed×regular															
SUCRE	78.0	29.0	42.3	29.4	10.9	15.9	62.0	19.5	**29.7**	59.1	6.5	11.7	46.9	46.9	**46.9**
UBIU	41.5	29.9	34.7	20.5	14.6	**17.0**	6.7	11.0	8.3	13.3	23.4	**17.0**	50.0	52.4	32.3
open×gold															
open×regular															
English															
closed×gold															
RelaxCor	100	100	100	75.6	75.6	**75.6**	21.9	72.4	33.7	74.8	97.0	**84.5**	57.0	83.4	61.3
SUCRE	100	100	100	74.3	74.3	74.3	68.1	54.9	**60.8**	86.7	78.5	82.4	77.3	67.0	**70.8**
TANL-1	99.8	81.7	89.8	75.0	61.4	67.6	23.7	24.4	24.0	74.6	72.1	73.4	51.8	68.8	52.1
UBIU	92.5	99.5	95.9	63.4	68.2	65.7	17.2	25.5	20.5	67.8	83.5	74.8	52.6	60.8	54.0
closed×regular															
SUCRE	78.4	83.0	80.7	61.0	64.5	**62.7**	57.7	48.1	**52.5**	68.3	65.9	**67.1**	58.9	65.7	**61.2**
TANL-1	79.6	68.9	73.9	61.7	53.4	57.3	23.8	25.5	24.6	62.1	60.5	61.3	50.9	68.0	49.3
UBIU	66.7	83.6	74.2	48.2	60.4	53.6	11.6	18.4	14.2	50.9	69.2	58.7	50.9	56.3	51.0
open×gold															
Corry-B	100	100	100	77.5	77.5	77.5	56.1	57.5	56.8	82.6	85.7	84.1	69.3	75.3	**71.8**
Corry-C	100	100	100	77.7	77.7	**77.7**	57.4	58.3	57.9	83.1	84.7	83.9	71.3	71.6	71.5
Corry-M	100	100	100	73.8	73.8	73.8	62.5	56.2	**59.2**	85.5	78.6	81.9	76.2	58.8	62.7
RelaxCor	100	100	100	75.8	75.8	75.8	22.6	70.5	34.2	75.2	96.7	**84.6**	58.0	83.8	62.7
open×regular															
BART	76.1	69.8	72.8	70.1	64.3	67.1	62.8	52.4	57.1	74.9	67.7	71.1	55.3	73.2	57.7
Corry-B	79.8	76.4	78.1	70.4	67.4	68.9	55.0	54.2	54.6	73.7	74.1	**73.9**	57.1	75.7	**60.6**
Corry-C	79.8	76.4	78.1	70.9	67.9	**69.4**	54.7	55.5	55.1	73.8	73.1	73.5	57.4	63.8	59.4
Corry-M	79.8	76.4	78.1	66.3	63.5	64.8	61.5	53.4	**57.2**	76.8	66.5	71.3	58.5	56.2	57.1
German															
closed×gold															
SUCRE	100	100	100	72.9	72.9	72.9	74.4	48.1	**58.4**	90.4	73.6	81.1	78.2	61.8	**66.4**
TANL-1	100	100	100	77.7	77.7	**77.7**	16.4	60.6	25.9	77.2	96.7	**85.9**	54.4	75.1	57.4
UBIU	92.6	95.5	94.0	67.4	68.9	68.2	22.1	21.7	21.9	73.7	77.9	75.7	60.0	77.2	64.5
closed×regular															
SUCRE	79.3	77.5	78.4	60.6	59.2	**59.9**	49.3	35.0	**40.9**	69.1	60.1	**64.3**	52.7	59.3	**53.6**
TANL-1	60.9	57.7	59.2	50.9	48.2	49.5	10.2	31.5	15.4	47.2	54.9	50.7	50.2	63.0	44.7
UBIU	50.6	66.8	57.6	39.4	51.9	44.8	9.5	11.4	10.4	41.2	53.7	46.6	50.2	54.4	48.0
open×gold															
BART	94.3	93.7	94.0	67.1	66.7	**66.9**	70.5	40.1	**51.1**	85.3	64.4	**73.4**	65.5	61.0	**62.8**
open×regular															
BART	82.5	82.3	82.4	61.4	61.2	**61.3**	61.4	36.1	**45.5**	75.3	58.3	**65.7**	55.9	60.3	**57.3**
Italian															
closed×gold															
SUCRE	98.4	98.4	98.4	66.0	66.0	**66.0**	48.1	42.3	**45.0**	76.7	76.9	**76.8**	54.8	63.5	**56.9**
closed×regular															
SUCRE	84.6	98.1	90.8	57.1	66.2	**61.3**	50.1	50.7	**50.4**	63.6	79.2	**70.6**	55.2	68.3	**57.7**
UBIU	46.8	35.9	40.6	37.9	29.0	32.9	2.9	4.6	3.6	38.4	31.9	34.8	50.0	46.6	37.2
open×gold															
open×regular															
BART	42.8	80.7	55.9	35.0	66.1	45.8	35.3	54.0	**42.7**	34.6	70.6	46.4	57.1	68.1	**59.6**
TANL-1	90.5	73.8	81.3	62.2	50.7	**55.9**	37.2	28.3	32.1	66.8	56.5	**61.2**	50.7	69.3	48.5
Spanish															
closed×gold															
RelaxCor	100	100	100	66.6	66.6	66.6	14.8	73.8	24.7	65.3	97.5	**78.2**	53.4	81.8	55.6
SUCRE	100	100	100	69.8	69.8	**69.8**	52.7	58.3	**55.3**	75.8	79.0	77.4	67.3	62.5	**64.5**
TANL-1	100	96.8	98.4	66.9	64.7	65.8	16.6	56.5	25.7	65.2	93.4	76.8	52.5	79.0	54.1
UBIU	73.8	96.4	83.6	45.7	59.6	51.7	9.6	18.8	12.7	46.8	77.1	58.3	52.9	63.9	54.3
closed×regular															
SUCRE	74.9	66.3	70.3	56.3	49.9	52.9	35.8	36.8	**36.3**	56.6	54.6	55.6	52.1	61.2	**51.4**
TANL-1	82.2	84.1	83.1	58.6	60.0	**59.3**	14.0	48.4	21.7	56.6	79.0	**66.0**	51.4	74.7	**51.4**
UBIU	51.1	72.7	60.0	33.6	47.6	39.4	7.6	14.4	10.0	32.8	57.1	41.6	50.4	54.6	48.4
open×gold															
open×regular															

Table 6.5: Official results of the participating systems for all languages, settings, and metrics

6.6 Conclusions

This paper has introduced the main features of the SemEval-2010 task on coreference resolution. The goal of the task was to evaluate and compare automatic coreference resolution systems for six different languages in four evaluation settings and using four different metrics. This complex scenario aimed at providing insight into several aspects of coreference resolution, including portability across languages, relevance of linguistic information at different levels, and behavior of alternative scoring metrics.

The task attracted considerable attention from a number of researchers, but only six teams submitted their final results. Participating systems did not run their systems for all the languages and evaluation settings, thus making direct comparisons between them very difficult. Nonetheless, we were able to observe some interesting aspects from the empirical evaluation.

An important conclusion was the confirmation that different evaluation metrics provide different system rankings and the scores are not commensurate. Attention thus needs to be paid to coreference evaluation. The behavior and applicability of the scoring metrics requires further investigation in order to guarantee a fair evaluation when comparing systems in the future. We hope to have the opportunity to thoroughly discuss this and the rest of interesting questions raised by the task during the SemEval workshop at ACL 2010.

An additional valuable benefit is the set of resources developed throughout the task. As task organizers, we intend to facilitate the sharing of datasets, scorers, and documentation by keeping them available for future research use. We believe that these resources will help to set future benchmarks for the research community and will contribute positively to the progress of the state of the art in coreference resolution. We will maintain and update the task website with post-SemEval contributions.

Acknowledgments We would like to thank the following people who contributed to the preparation of the task datasets: Manuel Bertran (UB), Oriol Borrega (UB), Orphée De Clercq (U. Ghent), Francesca Delogu (U. Trento), Jesús Giménez (UPC), Eduard Hovy (ISI-USC), Richard Johansson (U. Trento), Xavier Lluís (UPC), Montse Nofre (UB), Lluís Padró (UPC), Kepa Joseba Rodríguez (U. Trento), Mihai Surdeanu (Stanford), Olga Uryupina (U. Trento), Lente Van Leuven (UB), and Rita Zaragoza (UB). We would also like to thank LDC and die tageszeitung for distributing freely the English and German datasets.

This work was funded in part by the Spanish Ministry of Science and Innovation through the projects TEXT-MESS 2.0 (TIN2009-13391-C04-04), OpenMT-2 (TIN2009-14675-C03), and KNOW2 (TIN2009-14715-C04-04), and an FPU doctoral scholarship (AP2006-00994) held by M. Recasens. It also received financial support from the Seventh Framework Programme of the EU (FP7/2007-2013) under GA 247762 (FAUST), from the STEVIN program of the Nederlandse Taalunie through the COREA and SoNaR projects, and from the Provincia Autonoma di Trento through the LiveMemories project.

Part III

TEORIA DE LA COREFERÈNCIA

CAPÍTOL 7

On Paraphrase and Coreference

Marta Recasens and Marta Vila

University of Barcelona

To appear as a squib in *Computational Linguistics*, 36(4)

Abstract By providing a better understanding of paraphrase and coreference in terms of similarities and differences in their linguistic nature, this article delimits what the focus of paraphrase extraction and coreference resolution tasks should be, and to what extent they can help each other. We argue for the relevance of this discussion to Natural Language Processing.

7.1 Introduction

Paraphrase extraction[1] and coreference resolution have applications in Question Answering, Information Extraction, Machine Translation, and so forth. Paraphrase pairs might be coreferential, and coreference relations are sometimes paraphrases. The two overlap considerably (HIRST, 1981), but their definitions make them significantly different in essence: Paraphrasing concerns meaning, whereas coreference is about discourse referents. Thus, they do not always coincide. In the following example, *b* and *d* are both coreferent and paraphrastic, whereas *a*, *c*, *e*, *f*, and *h* are coreferent but not paraphrastic, and *g* and *i* are paraphrastic but not coreferent.

[1] Recognition, extraction, and generation are all paraphrase-related tasks. We will center ourselves on paraphrase extraction, as this is the task in which paraphrase and coreference resolution mainly overlap.

(1) [Tony]$_a$ went to see [the ophthalmologist]$_b$ and got [his]$_c$ eyes checked. [The eye doctor]$_d$ told [him]$_e$ that [his]$_f$ [cataracts]$_g$ were getting worse. [His]$_h$ mother also suffered from [cloudy vision]$_i$.

The discourse model built for Example (1) contains six entities (i.e., Tony, the eye doctor, Tony's eyes, Tony's cataracts, Tony's mother, cataracts). Because *a*, *c*, *e*, *f* and *h* all point to Tony, we say that they are coreferent. In contrast, in paraphrasing, we do not need to build a discourse entity to state that *g* and *i* are paraphrase pairs; we restrict ourselves to semantic content and this is why we check for sameness of meaning between *cataracts* and *cloudy vision* alone, regardless of whether they are a referential unit in a discourse. Despite the differences, it is possible for paraphrasing and coreference to co-occur, as in the case of *b* and *d*.

NLP components dealing with paraphrasing and coreference seem to have great potential to improve understanding and generation systems. As a result, they have been the focus of a large amount of work in the past couple of decades (see the surveys by Androutsopoulos and Malakasiotis [2010], Madnani and Dorr [2010], Ng [2010], and Poesio et al. [forthcoming]). Before computational linguistics, coreference had not been studied on its own from a purely linguistic perspective but was indirectly mentioned in the study of pronouns. Although there have been some linguistic works that consider paraphrasing, they do not fully respond to the needs of paraphrasing from a computational perspective.

This article discusses the similarities between paraphrase and coreference in order to point out the distinguishing factors that make paraphrase extraction and coreference resolution two separate yet related tasks. This is illustrated with examples extracted/adapted from different sources (DRAS, 1999; DODDINGTON *et al.*, 2004; DOLAN *et al.*, 2005; RECASENS i MARTÍ, 2010; VILA *et al.*, 2010) and our own. Apart from providing a better understanding of these tasks, we point out ways in which they can mutually benefit, which can shed light on future research.

7.2 Converging and diverging points

This section explores the overlapping relationship between paraphrase and coreference, highlighting the most relevant aspects that they have in common as well as those that distinguish them. They are both sameness relations (Section 7.2.2), but one is between meanings and the other between referents (Section 7.2.1). In terms of linguistic units, coreference is mainly restricted to noun phrases (NPs), whereas paraphrasing goes beyond and includes word-, phrase- and sentence-level expressions (Section 7.2.3). One final diverging point is the role they (might) play in discourse (Section 7.2.4).

7.2.1 Meaning and reference

The two dimensions that are the focus of paraphrasing and coreference are meaning and reference, respectively. Traditionally, paraphrase is defined as the relation

		Paraphrase ✔	Paraphrase ✗
Coreference	✔	(1,1) Tony went to see *the ophthalmologist* and got his eyes checked. *The eye doctor* told him ...	(1,2) *Tony* went to see the ophthalmologist and got *his* eyes checked.
	✗	(2,1) *ophthalmologist* *eye doctor*	(2,2) *His cataracts* were getting worse. *His mother* also suffered from cloudy vision.

Table 7.1: Paraphrase–coreference matrix

between two expressions that have the same *meaning* (i.e., they evoke the same mental concept), whereas coreference is defined as the relation between two expressions that have the same *referent* in the discourse (i.e., they point to the same entity). We follow KARTTUNEN (1976) and talk of "discourse referents" instead of "real-world referents."

In Table 7.1, the italicized pairs in cells (1,1) and (2,1) are both paraphrastic but they only corefer in (1,1). We cannot decide on (non-)coreference in (2,1) as we need a discourse to first assign a referent. In contrast, we can make paraphrasing judgments without taking discourse into consideration. Pairs like the one in cell (1,2) are only coreferent but not paraphrases because the proper noun *Tony* and the pronoun *his* have reference but no meaning. Lastly, neither phenomenon is observed in cell (2,2).

7.2.2 Sameness

Paraphrasing and coreference are usually defined as sameness relations: Two expressions that have the *same meaning* are paraphrastic, and two expressions that refer to the *same entity* in a discourse are coreferent. The concept of *sameness* is usually taken for granted and left unexplained, but establishing sameness is not straightforward. A strict interpretation of the concept makes sameness relations only possible in logic and mathematics, whereas a sloppy interpretation makes the definition too vague. In paraphrasing, if the loss of *at the city* in Example (2-b) is not considered to be relevant, Examples (2-a) and (2-b) are paraphrases; but if it is considered to be relevant, then they are not. It depends on where we draw the boundaries of what is accepted as the "same" meaning.

(2) a. The waterlogged conditions that ruled out play yesterday still prevailed *at the city* this morning.
b. The waterlogged conditions that ruled out play yesterday still prevailed this morning.

(3) On homecoming night *Postville* feels like Hometown, USA ... For those who prefer *the old Postville*, Mayor John Hyman has a simple answer.

Similarly, with respect to coreference, whether *Postville* and *the old Postville* in Example (1-c) are or are not the same entity depends on the granularity of the discourse. On a sloppy reading, one can assume that because Postville refers to the same spatial coordinates, it is the same town. On a strict reading, in contrast, drawing a distinction between the town as it was at two different moments in time results in two different entities: the old Postville versus the present-day Postville. They are not the same in that features have changed from the former to the latter.

The concept of sameness in paraphrasing has been questioned on many occasions. If we understood "same meaning" in the strictest sense, a large number of paraphrases would be ruled out. Thus, some authors argue for a looser definition of paraphrasing. BHAGAT (2009), for instance, talks about "quasi-paraphrases" as "sentences or phrases that convey approximately the same meaning." MILIĆEVIĆ (2007) draws a distinction between "exact" and "approximate" paraphrases. Finally, FUCHS (1994) prefers to use the notion of "equivalence" to "identity" on the grounds that the former allows for the existence of some semantic differences between the paraphrase pairs. The concept of identity in coreference, however, has hardly been questioned, as prototypical examples appear to be straightforward (e.g., *Barack Obama* and *Obama* and *he*). Only recently have RECASENS *et al.* (2010*a*) pointed out the need for talking about "near-identity" relations in order to account for cases such as Example (3), proposing a typology of such relations.

7.2.3 Linguistic units

Another axis of comparison between paraphrase and coreference concerns the types of linguistic units involved in each relation. Paraphrase can hold between different linguistic units, from morphemes to full texts, although the most attention has been paid to word-level paraphrase (*kid* and *child* in Example (4)), phrase-level paraphrase (*cried* and *burst into tears* in Example (4)), and sentence-level paraphrase (the two sentences in Example (4)).

(4) a. The kid cried.
b. The child burst into tears.

In contrast, coreference is more restricted in that the majority of relations occur at the phrasal level, especially between NPs. This explains why this has been the largest focus so far, although prepositional and adverbial phrases are also possible yet less frequent, as well as clauses or sentences. Coreference relations occur indistinctively between pronouns, proper nouns, and full NPs that are *referential*, namely, that have discourse referents. For this reason, pleonastic pronouns, nominal predicates, and appositives cannot enter into coreference relations. The first do not refer to any entity but are syntactically required; the last two express properties of an entity rather than introduce a new one. But this is an issue ignored

by the corpora annotated for the MUC and ACE programs (HIRSCHMAN i CHINCHOR, 1997; DODDINGTON *et al.*, 2004), hence the criticism by VAN DEEMTER i KIBBLE (2000).

In the case of paraphrasing, it is linguistic expressions that lack meaning (i.e., pronouns and proper nouns) that should not be treated as members of a paraphrase pair on their own (Example (5-a)) because paraphrase is only possible between meaningful units. This issue, however, takes on another dimension when seen at the sentence level. The sentences in Example (5-b) can be said to be paraphrases because they themselves contain the antecedent of the pronouns *I* and *he*.

(5) a. (i) A. Jiménez
(ii) I
b. (i) The Atlético de Madrid goalkeeper, A. Jiménez, yesterday realized one of his dreams by defeating Barcelona: "I had never beaten Barcelona."
(ii) The Atlético de Madrid goalkeeper, A. Jiménez, yesterday realized one of his dreams by defeating Barcelona, and said that he had never beaten Barcelona.

In Example (5-b), *A. Jiménez* and *I*/*he* continue not being paraphrastic in isolation. Polysemic, underspecified and metaphoric words show a slightly different behavior. It is not possible to establish paraphrase between them when they are deprived of context (CALLISON-BURCH, 2007, Chapter 4). In Example (6-a), *police officers* could be patrol police officers, and *investigators* could be university researchers. However, once they are embedded in a disambiguating context that fills them semantically, as in Example (6-b), then paraphrase can be established between *police officers* and *investigators*.

(6) a. (i) Police officers
(ii) Investigators
b. (i) *Police officers* searched 11 stores in Barcelona.
(ii) The *investigators* conducted numerous interviews with the victim.

As a final remark, and in accordance with the approach by FUCHS (1994), we consider Example (7)-like paraphrases that FUJITA (2005) and MILIĆEVIĆ (2007) call, respectively, "referential" and "cognitive" to be best treated as coreference rather than paraphrase, because they only rely on referential identity in a discourse.

(7) a. They got married *last year*.
b. They got married *in 2004*.

7.2.4 Discourse function

A further difference between paraphrasing and coreference concerns their degree of dependency on discourse. Given that coreference establishes sameness rela-

tions between the entities that populate a discourse (i.e., discourse referents), it is a linguistic phenomenon whose dependency on discourse is much stronger than paraphrasing. Thus, the latter can be approached from a discursive or a non-discursive perspective, which in turn allows for a distinction between reformulative paraphrasing (Example (8)) and non-reformulative paraphrasing (Example (9)).

(8) Speaker 1: Then they also diagnosed *a hemolytic–uremic syndrome*.
Speaker 2: What's that?
Speaker 1: *Renal insufficiency, in the kidneys*.

(9) a. X wrote Y.
b. X is the author of Y.

Reformulative paraphrasing occurs in a reformulation context when a rewording of a previously expressed content is added for discursive reasons, such as emphasis, correction or clarification. Non-reformulative paraphrasing does not consider the role that paraphrasing plays in discourse. Reformulative paraphrase pairs have to be extracted from a single piece of discourse; non-reformulative paraphrase pairs can be extracted—each member of the pair on its own—from different discourse pieces. The reformulation in the third utterance in Example (8) gives an explanation in a language less technical than that in the first utterance; whereas Example (9-a) and Example (9-b) are simply two alternative ways of expressing an authorship relation.

The strong discourse dependency of coreference explains the major role it plays in terms of cohesion. Being such a cohesive device, it follows that intra-document coreference, which takes place within a single discourse unit (or across a collection of documents linked by topic), is the most primary. Cross-document coreference, on the other hand, constitutes a task on its own in NLP but falls beyond the scope of linguistic coreference due to the lack of a common universe of discourse. The assumption behind cross-document coreference is that there is an underlying global discourse that enables various documents to be treated as a single macro-document.

Despite the differences, the discourse function of reformulative paraphrasing brings it close to coreference in the sense that they both contribute to the cohesion and development of discourse.

7.3 Mutual benefits

Both paraphrase extraction and coreference resolution are complex tasks far from being solved at present, and we believe that there could be improvements in performance if researchers on each side paid attention to the others. The similarities (i.e., relations of sameness, relations between NPs) allow for mutual collaboration, whereas the differences (i.e., focus on either meaning or reference) allow for resorting to either paraphrase or coreference to solve the other. In general, the greatest benefits come for cases in which either paraphrase or coreference are especially difficult to detect automatically. More specifically, we see direct mutual benefits

when both phenomena occur either in the same expression or in neighboring expressions.

For pairs of linguistic expressions that show both relations, we can hypothesize paraphrasing relationships between NPs for which coreference is easier to detect. For instance, coreference between the two NPs in Example (10) is very likely given that they have the same head, head match being one of the most successful features in coreference resolution (HAGHIGHI i KLEIN, 2009). In contrast, deciding on paraphrase would be hard due to the difficulty of matching the modifiers of the two NPs.

(10) a. The director of a multinational with huge profits.
b. The director of a solvent company with headquarters in many countries.

In the opposite direction, we can hypothesize coreference links between NPs for which paraphrasing can be recognized with considerable ease (Example (11)). Light elements (e.g., *fact*), for instance, are normally taken into account in paraphrasing—but not in coreference resolution—as their addition or deletion does not involve a significant change in meaning.

(11) a. The creation of a company.
b. The fact of creating a company.

By neighboring expressions, we mean two parallel structures each containing a coreferent mention of the same entity next to a member of the same paraphrase pair. Note that the coreferent expressions in the following examples are printed in *italics* and the paraphrase units are printed in **bold**. If a resolution module identifies the coreferent pairs in Example (12), then these can function as two anchor points, *X* and *Y*, to infer that the text between them is paraphrastic: *X complained today before Y*, and *X is formulating the corresponding complaint to Y*.

(12) a. *Argentina*$_X$ **complained today before** *the British Government*$_Y$ about the violation of the air space of this South American country.
b. *This Chancellorship*$_X$ **is formulating the corresponding complaint to** *the British Government*$_Y$ for this violation of the Argentinian air space.

Some authors have already used coreference resolution in their paraphrasing systems in a similar way to the examples herein. SHINYAMA i SEKINE (2003) benefit from the fact that a single event can be reported in more than one newspaper article in different ways, keeping certain kinds of NPs such as names, dates, and numbers unchanged. Thus, these can behave as anchor points for paraphrase extraction. Their system uses coreference resolution to find anchors which refer to the same entity.

Conversely, knowing that a stretch of text next to an NP paraphrases another stretch of text next to another NP helps to identify a coreference link between

the two NPs, as shown by Example (13), where two diction verbs are easily detected as a paraphrase and thus their subjects can be hypothesized to corefer. If the paraphrase system identifies the mapping between the indirect speech in Example (13-a) and the direct speech in Example (13-b), the coreference relation between the subjects is corroborated. Another difficult coreference link that can be detected with the help of paraphrasing is Example (14): If the predicates are recognized as paraphrases, then the subjects are likely to corefer.

(13) a. *The trainer of the Cuban athlete Sotomayor* **said** that the world record holder is in a fit state to win the Games in Sydney.
b. "The record holder is in a fit state to win the Olympic Games," **explained** *De la Torre*.

(14) a. *Police officers* **searched 11 stores in Barcelona**.
b. *The investigators* **carried out 11 searches in stores in the center of Barcelona**.

Taking this idea one step further, new coreference resolution strategies can be developed with the aid of shallow paraphrasing techniques. A two-step process for coreference resolution might consist of hypothesizing first sentence-level paraphrases via *n*-gram or named-entity overlapping, aligning phrases that are (possible) paraphrases, and hypothesizing that they corefer. Second, a coreference module can act as a filter and provide a second classification. Such a procedure could be successful for the cases exemplified in Examples (12) to (14).

This strategy reverses the tacit assumption that coreference is solved before sentence-level paraphrasing. Meaning alone does not make it possible to state that the two pairs in Example (5-b), repeated in Example (15), or the two pairs in Example (16) are paraphrases without first solving the coreference relations.

(15) a. *The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona: "*I* had never beaten Barcelona."
b. *The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona, and said that *he* had never beaten Barcelona.

(16) a. Secretary of State Colin Powell last week ruled out *a non-aggression treaty*.
b. But Secretary of State Colin Powell brushed off *this possibility*.

However, cooperative work between paraphrasing and coreference is not always possible, and it is harder if neither of the two can be detected by means of widely used strategies. In other cases, cooperation can even be misleading. In Example (17), the two bold phrases are paraphrases, but their subjects do not corefer. The detection of words like *another* (Example (17-b)) gives a key to help to prevent this kind of error.

(17) a. A total of 26 Cuban citizens remain in the police station of the airport of Barajas **after requesting political asylum**.
b. Another three Cubans **requested political asylum**.

On the basis of these various examples, we claim that a full understanding of both the similarities and disparities will enable fruitful collaboration between researchers working on paraphrasing and those working on coreference. Even more importantly, our main claim is that such an understanding about the fundamental linguistic issues is a prerequisite for building paraphrase and coreference systems not lacking in linguistic rigor. In brief, we call for the return of linguistics to paraphrasing and coreference automatic applications, as well as to NLP in general, adhering to the call by WINTNER (2009:643), who cites examples that demonstrate "what computational linguistics can achieve when it is backed up and informed by linguistic theory."

Acknowledgements We are grateful to Eduard Hovy, M. Antònia Martí, Horacio Rodríguez, and Mariona Taulé for their helpful advice as experienced researchers. We would also like to express our gratitude to the three anonymous reviewers for their suggestions to improve this article.

This work was partly supported by FPU Grants AP2006-00994 and AP2008-02185 from the Spanish Ministry of Education, and Project TEXT-MESS 2.0 (TIN2009-13391-C04-04).

CAPÍTOL 8

Identity, Non-identity, and Near-identity: Addressing the complexity of coreference

Recasens, Marta*, Eduard Hovy**, and M. Antònia Martí*

*University of Barcelona
**USC Information Sciences Institute

Submitted to *Lingua*

Abstract This article examines the mainstream categorical definition of coreference as 'identity of the real-world referents.' It argues that coreference is best handled when identity is treated as a continuum, ranging from full identity to non-identity, with room for near-identity relations to explain currently problematic cases. This middle ground is needed because in real text, linguistic expressions often stand in relations that are neither full coreference nor non-coreference, a situation that has led to contradictory treatment of cases in previous coreference annotation efforts. We discuss key issues for coreference such as conceptual categorization, individuation, criteria of identity, and the discourse model construct. We define coreference as a scalar relation between two (or more) linguistic expressions that refer to discourse entities considered to be at the same granularity level relevant to the pragmatic purpose. We present a typology of coreference relations, including various types of near-identity, that is developed and validated in a series of annotation exercises. We describe the operation of the coreference relations in terms of Fauconnier's mental space theory.

Keywords Coreference · Discourse · Categorization · Near-identity · Specification · Refocusing · Neutralization

8.1 Introduction

Coreference phenomena have been treated by theoretical linguists who study the relation between pronouns or definite descriptions and their antecedents, by discourse analysts who research factors contributing to coherence, by psycholinguists interested in the knowledge intervening in the interpretation of coreferent expressions, by logicians and language philosophers who analyze propositions in terms of existence and truth conditions, and by computational linguists who attempt to build coreference resolution systems that automatically identify coreferent expressions in a text. Despite the varied interests, common to all them is the understanding of coreference as 'identity of reference,' namely a relation holding between linguistic expressions that refer to the same entity. This apparently straightforward definition, however, hides a number of unexamined assumptions about reference and identity that we set out to explore in this article.

The shortcomings of the current definition become especially apparent when real corpora are annotated with coreference information (VERSLEY, 2008; POESIO i ARTSTEIN, 2005), since the low levels of inter-annotator agreement usually obtained seem to go against the simplicity of the definition. Compare the two annotations for (1) and (2), where coreferent noun phrases (NPs) are printed in italics, and (a) and (b) are drawn from the ACE (DODDINGTON *et al.*, 2004) and OntoNotes (PRADHAN *et al.*, 2007*a*) corpora, respectively.

(1) a. On homecoming night *Postville* feels like Hometown, USA, but a look around *this town of 2,000* shows *it*'s become a miniature Ellis Island. *This* was an all-white, all-Christian community ... For those who prefer the old Postville, Mayor John Hyman has a simple answer.

b. On homecoming night *Postville* feels like Hometown, USA, but a look around *this town of 2,000* shows *it*'s become a miniature Ellis Island. *This* was an all-white, all-Christian community ... For those who prefer *the old Postville*, Mayor John Hyman has a simple answer.

(2) a. Last night in Tel Aviv, *Jews* attacked a restaurant that employs Palestinians. "We want war," *the crowd* chanted.

b. Last night in Tel Aviv, Jews attacked a restaurant that employs Palestinians. "*We* want war," *the crowd* chanted.

The complexity exemplified by (1) and (2) arises when two references denote 'almost' the same thing, either for a single individual—*Postville* and *the old Postville* (1)—or across two groups—*Jews*, *we*, and *the crowd* (2). Such cases are indicative that the predominant categorical distinction between coreference (identity) and non-coreference (non-identity) is too limited—assuming that categorization in discourse is a pre-fixed process instead of a dynamic one—and so fails when confronted with the full range of natural language phenomena. Rather, coreference is best viewed as a continuum ranging from identity to non-identity, with room for near-identity relations to handle currently problematic cases that do not fall neatly

into either full coreference or non-coreference.

The goal of this article is to develop a richer, more detailed understanding of coreference phenomena that explains under what circumstances linguistic expressions are interpreted as coreferent, or quasi-coreferent. To this end, we propose a novel framework that draws on insights from JACKENDOFF (1983, 2002), FAUCONNIER (1985, 1997), GEACH (1967), HOBBS (1985), NUNBERG (1984) and BARKER (2010) among others. The framework resulted from reviewing key issues at the basis of coreference such as conceptual categorization, individuation, criteria of identity, and the role of pragmatics. In brief, we redefine coreference as a scalar relation between discourse entities (DEs, henceforth) conceived of as the same at the granularity level relevant to the pragmatic purpose. This leads us to propose a continuum from identity to non-identity through near-identity, which occurs when entities share most but not all feature values. We present a typology of those features whose change elicits a near-identity relation, which we account for in terms of three cognitive operations of categorization: specification, refocusing and neutralization. The former two create new indexical features, while the latter neutralizes potential indexical features. Such an understanding has consequences for the various branches of linguistics, from theoretical to psycho- and computational linguistics.

8.2 Background

Since coreference touches on subjects such as reference, categorization, and identity about which an extensive philosophical and linguistic literature exists, we can partly build on previous research. Only partly, however, because, as this section will reveal, there is a gap between real data and much previous theoretical work—which mostly uses prefabricated examples—that makes it unable to account for the problems exhibited by naturally occurring data.[1] In this section, we discuss the main drawbacks of existing accounts while reviewing the main ideas from previous work that are relevant to our account of coreference, which will be fully presented in the next section. Throughout we make explicit the assumptions and commitments underlying our approach. At the risk of getting into deeply philosophical discussions, we will limit ourselves to the key ideas that serve as the basis to develop our coreference framework.

In order to make it easier for the reader to follow the thread of this section, Fig. 8.1 is a concise diagram that connects the topics we will discuss. The shaded ovals indicate the relevant sections in this paper. Inside the box, the bottom sequence should be understood as one of the dimensions contained by the top sequence, i.e., language as part of our cognitive apparatus. We will start by defining the projected world in opposition to what we call 'the world' to then explore the elements and processes involved in the construction of the projected world. En-

[1] Appendix B includes the kind of real data that traditional models have failed to explain and that we build on for this article.

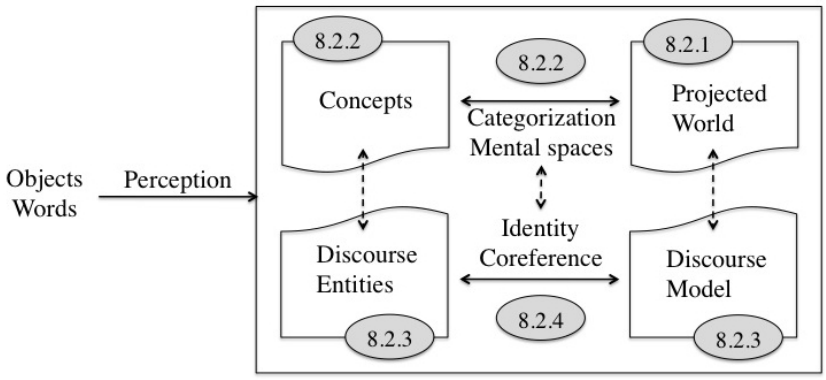

Figure 8.1: Language and cognition

tering the domain of language, we will consider the language-specific counterparts to concepts—i.e., DEs—and to the projected world—i.e., the discourse model. Finally, we will get to our main subject of interest: identity relations and coreference, which play a key role in organizing DEs in the discourse model.

8.2.1 What reference is about

The realist theory that views reference as about the real world has underlain traditional theories of meaning from the theory of mediated reference (FREGE, 1892), where a distinction is drawn between *sense* (intension) and *reference* (extension), to the theory of direct reference (RUSSELL, 1905), where meaning is equated with reference. Common to them is the assumption that the target of linguistic reference is the objective, *real world*, whether directly or mediated by a sense. It was not until the advent of cognitive semantics in the 1980s that this view began to be questioned in semantics.[2] Drawing upon empirical findings from the school of Gestalt psychology, JACKENDOFF (1983) argues for a conceptualist theory of reference according to which the information conveyed by language refers to entities in the world as conceptualized by the language user. This world he calls it the *projected world*. Since a number of mental processes participate in processing our environmental input, "one cannot perceive the real world as it is." Rather, the projected world is the world as we experience it, i.e., constructed by our perceptual systems in response to whatever is "out there."

Following JACKENDOFF (1983), we need to distinguish between the real world as the source of environmental input and the projected world as the experienced world. In fact, the study of language does not need to take the real world into account but only the projected world, as direct access to the former is barred to us and so our linguistic expressions must necessarily refer to the latter. An immediate corollary is that language is necessarily subjective. That does not however imply

[2]Before, in the 18th century, the philosopher Kant had distinguished the *noumenal world* (the real world) from the *phenomenal world* (the world we perceive).

unprincipled variability. The fact that the processes by which we construct the projected world are universal makes our projections compatible to a major extent, thus enabling communication.

LAKOFF's (1987) *experientialism*, while acknowledging the existence of the real world, is also based on the idea that all our perceptions are filtered by the body and so we cannot access any but the world as processed by our cognitive apparatus. He emphasizes the crucial consequences that the embodiment of thought entails, i.e., our understanding of the world is largely determined by the form of the human body. From this perspective, recurring patterns of understanding and reasoning such as metaphor and metonymic mappings that condition our perceptions of the world are in turn formed from our bodily interactions.

By dissociating our account of coreference from real-world referents, we can abandon the real world and thus the requirements imposed by identity judgments in terms of an objective, unique, world that often result in dead-end contradictions. Instead, the way entities are built in language is closely tied to our cognitive apparatus rather than to intrinsic properties of the entities themselves. The discourse model parallels the projected world.

8.2.2 Categorizing the projected world

Once we have replaced the real world with the projected world, we need to consider what forms and provides structure and regular behavior to the projected world, which immediately brings us to mental information, conceptual structures, categories, and the like, and at this point we start treading on thin ice for much remains unknown when it comes to the brain. As we will see in Section 8.3, the theories that most help explain the coreference facts come from JACKENDOFF's (1983) conceptual semantics and FAUCONNIER's (1985) mental space theory.

Concepts and categories are closely intertwined, the former referring to all the knowledge that one has about categories—collections of instances which are treated as if they were the same. By arguing against the classical Aristotelian view that categories are defined by necessary and sufficient conditions—WITTGENSTEIN (1953) being a precedent—JACKENDOFF (1983) claims that categories in the projected world are determined by complex perceptual and cognitive principles. Entities are not given by the external physical world, but it is the human cognitive apparatus that carves up the projected world into seemingly distinct and distinguishable categories, thus making divisions where there are none in the world.

JACKENDOFF (1983) argues that for an entity to be projected there must be a corresponding conceptual constituent. We *construct* entities from the environmental input according to the concepts that we have experienced, learned, and structured in terms of prototypes and basic-level concepts (LAKOFF, 1987). The situation itself, our previous experience, our intentions or needs, can make certain features more salient than others and lead us towards a particular individuation. A very important point in the categorization process is that it is graded rather than categorical. We are born with an "ability to conceptualize the world at different

granularities and to switch among these granularities" (HOBBS, 1985). Thus, a *mountain* is categorized differently depending on the situation: we will think of it as a very large hill when talking to a child; as a steep slope when going skiing; or as a volume that can be excavated when doing geology. Taking this flexibility into account will be key to understand how coreference works.

FAUCONNIER's (1985; 1997) mental space theory is especially interesting for the present work as it was originally developed to address problems of indirect reference and referential opacity, although it has become useful to explain language phenomena and cognition in general. To this end, for purposes of local understanding it provides a frame of abstract mental structures known as *mental spaces* that are constructed while we think and talk, and in which referential structure can be projected. Mental spaces organize the unconscious processes that take place behind the scenes and that are highly responsible for meaning construction. The details of how they are set up and configured will become evident in Section 8.3.2.

8.2.3 Building DEs

It is by connecting to conceptual structures that language acquires meaning, and there can be no reference without conceptualization: "A language user cannot refer to an entity without having some conceptualization of it" (JACKENDOFF, 2002). Note, however, that being in the real world is not a necessary condition for reference, and an entity's being in the world is not sufficient for reference either. The crucial feature for linguistic reference is to have what JACKENDOFF (2002) calls an *indexical feature* established by our mind in response to a perceptual input. An indexical feature brings about the construction of a *discourse referent* (KARTTUNEN, 1976) or a DE (WEBBER, 1979). These are the instances we talk about by means of referring expressions, believing that they are objects "out there."

As a discourse evolves, DEs grow in number and populate the *discourse model*, which is a temporary mental 'toy' replica of the projected world built by language users specifically for interpreting a particular discourse. Thus, categorization in discourse occurs dynamically rather than statically. Apart from the collection of DEs, the discourse model includes the information that is said about them, i.e., their properties and the relations they participate in, and this information accumulates as the discourse progresses. Properties may validly be changed or introduced in the discourse that are clearly untrue of the original 'real-world' referents. Coreference relations occur thus not between 'actual' referents but between DEs. Like any other construct, DEs are subjective in that they ultimately depend on a language user's specific discourse model. However, as is the case with the projected world, there tends to be a high degree of similarity between the discourse models built by different language users, at least within the same culture, and within the same discourse. This notwithstanding, misunderstandings might be caused by relevant differences between different models.

Various formal representations of the discourse model have been suggested such as KAMP's (1981) Discourse Representation Theory or HEIM's (1983) File

Change Semantics. According to the view of language for which we argue (Fig. 8.1), these are too restricted to the language level and largely ignore general cognitive processes that are not language-specific. In contrast, a more flexible representation integrating both language and cognition is provided by mental spaces (FAUCONNIER, 1985).[3]

A final point to be made in relation to DEs concerns their ontological type. Each type has its own characteristic conditions of identity and individuation, which has consequences on coreference decisions. FRAURUD (1996) links ontological type with form of referring expression and suggests three main types. The most individuated entities are *Individuals*, i.e., "entities that are conceived of in their own right and that are directly identifiable, generally by means of a proper name." A proper name has an indexical in its associated concept. Individuals are opposed on the one hand to *Functionals*, entities that "are conceived of only in relation to other entities" (e.g., a person's nose), and on the other hand to *Instances*, entities that "are merely conceived of as instantiations of types," such as a glass of wine. A different kind of knowledge is involved in interpreting each class: token or referent knowledge for Individuals; relational type knowledge for Functionals; and sortal type knowledge for Instances. However, whether a certain entity is conceived of as one or other class is not a categorical question, but a matter of degree of individuation—or granularity.

8.2.4 Identity in the discourse model

Identity judgments between DEs become coreference judgments. As already hinted, we view coreference as the relation between expressions that refer to the same DE in the discourse model. Our approach to identity—and 'sameness'—lies within the domain of discourse and distances itself from logical or philosophical ones, where applying an *absolute* notion of identity to the ever-changing physical world results in a number of paradoxes (Theseus's Ship, Heraclitus' river, Chrysippus' Paradox, the Statue and the Clay, etc.).

As pointed out by FAUCONNIER (1997, pg. 51), "a natural-language sentence is a completely different kind of thing from a sentence in a logical calculus." Mathematical formulas give structural information explicitly and unambiguously. In contrast, language expressions do not have a meaning in itself but only a meaning *potential*. Thus, natural-language sentences are best seen as "a set of (underspecified) instructions for cognitive construction" that allow for producing meaning within a discourse and a context. The so-called Leibniz's Law[4] fails in opaque contexts as exemplified by (3), where James Bond, the top British spy, has been introduced to Ursula as Earl Grey, the wealthiest tea importer. If the wealthiest tea

[3] A preliminary application of mental space theory to complex coreference phenomena occurs in VERSLEY (2008).

[4] Leibniz's Law or the Principle of the Identity of Indiscernibles state, respectively, that,
For all x and y, if $x = y$, then x and y have the same properties.
For all x and y, if x and y have the same properties, then $x = y$.

importer is actually the very ugly Lord Lipton, then (3-a) is true, whereas (3-b) is false. Note that although the two names/descriptions are true of the same referent, one cannot be substituted for the other *salva veritate* due to their being embedded in Ursula's beliefs.

(3) a. Ursula thinks the wealthiest tea importer is handsome.
b. Ursula thinks Lord Lipton is handsome.

In response to the notion of absolute identity, GEACH (1967) argues that there is only *relative* identity.[5] An identity judgment must always be accompanied by some particular standard of sameness. That in accordance with which we judge corresponds to GEACH's (1962:39) *criterion of identity*, which he identifies as a common noun *A*—a sortal concept—typically understood from the context of utterance: "*x* is the same *A* as *y* but *x* and *y* are different *G*s." Reprising example (1) from Section 8.1, for which a notion of absolute identity produces two contradictory annotations, we find in GEACH's relative identity a satisfactory explanation: the old and the new Postville both refer to the 'same city' but to two different temporal instances: the city of Postville at time_1 (a white, Christian community) and the city of Postville at time_2 (with 2,000 citizens from varied nationalities).

This case exemplifies the general problem of change and identity, i.e., how identity is preserved over time, for which two major philosophical theories exist. Endurantism views entities as wholly present at every moment of their existence. On the other hand, perdurantism claims that entities are four dimensional—the fourth dimension being time—and that they have temporal parts. For perdurantists we can talk about entities not only in a temporal way (e.g., the old Postville versus the new Postville), but also in an atemporal way taking in all times at once (e.g., the city of Postville). In a similar vein, BARKER (2010) points out that some sentences are theoretically—but not pragmatically—ambiguous between two readings: an *individual-level* or type reading, and a *stage-level* or token reading. The former results in a hypo-individuation while the latter in a hyper-individuation. It is also from this perspective that the identity between 'coreferent' discourse referents that evolve through discourse is considered by CHAROLLES i SCHNEDECKER (1993).

The different granularity levels at which we categorize—and thus at which DEs can be construed—make it possible for us to conceive of identity relations at different degrees, more or less coarse. The degree of individuation is largely determined by the context of the discourse. In HOBBS's (1985) words, "we look at the world under various grain sizes and abstract from it only those things that serve our present interests." Linguistic studies that elaborate on the use of the words *same* and *different* (NUNBERG, 1984; BAKER, 2003; BARKER, 2010; LASERSOHN, 2000) coincide in that identity judgments take into consideration only those properties that are relevant to the pragmatic purpose, that is, "when we say that *a* and *b* are the same, we mean simply that they are the same for purposes of argument"

[5] We still believe, however, that absolute identity exists at least as a mental concept relative to which the more useful notion of relative identity is understood.

(NUNBERG, 1984:207).

In terms of FAUCONNIER's (1997) mental space theory, a sentence in itself has no fixed number of readings, and different space configurations result in different construals of DEs. The choice between the formally possible configurations is partly resolved by pragmatic considerations such as relevance, noncontradiction, prototypicality, etc. These should be taken into account in a full-fledged description of coreference as they have consequences in determining the criteria of identity used for establishing coreference relations. The range of identity criteria explains that coreference is best approached as a continuum.

8.2.5 Summary

We can conclude this section with the following major assumptions,

(i) There is no unique physical world to which referring expressions all point, but a host of individual worlds projected by our minds.

(ii) DEs are constructed based on the concepts and categories responsible for building the projected world, thus with the same potential range of individuation.

(iii) The discourse model is the mental space dynamically constructed for discourse understanding, and so is the space where coreference takes place.

(iv) Coreference relations between DEs depend on criteria of identity determined by the communicative purposes.

8.3 Coreference along a continuum

The different elements presented in the previous section are integrated here into a single framework with the aim of reducing the gap between theoretical claims and empirical data. We start by redefining coreference as it is currently understood, followed by a description of the mental space framework and formal notation. This will provide the tools to present our continuum model for coreference as well as the operations of specification, refocusing and neutralization that we use to account for coreference in real data.

8.3.1 Definition

The mainstream definition of coreference can be phrased as

> Coreference is a relation holding between two (or more) linguistic expressions that refer to the same entity in the real world.

This definition presents two major problems: its assumption that 'sameness' is a straightforward relation, and its commitment to the 'real world' as the domain of

entities to which language refers. We propose the following alternative definition that forms the basis of our coreference model:

> Coreference is a scalar relation holding between two (or more) linguistic expressions that refer to DEs considered to be at the same granularity level relevant to the pragmatic purpose.

Note that there are three keywords in this new definition. First, we no longer allude to the real world; rather, we place the coreference phenomenon within the discourse model, thus 'DEs.' Second, these entities are constructs of conceptualization mechanisms and, since there are degrees of individuation, the identity relation only holds at a certain 'granularity level.' Last, the granularity level is set at the value that is 'relevant' to the pragmatics of the particular discourse.

8.3.2 Fauconnier's mental spaces

The general structure of our framework draws on FAUCONNIER's (1985, 1997) mental space theory. Its value lies in the tools it provides for making explicit the construction of meaning from the (underspecified) forms of language, as these themselves contain little of what goes into meaning construction. By operating at the conceptual level and unlike truth-conditional approaches, mental spaces allow of a broad range of potential meanings that are narrowed down conveniently as a function of the discourse context. Our main two focuses will be mental space elements—high-order mental entities corresponding to DEs that are named by NPs—and the connections between them. Showing how connections are established, and how it affects coreference judgments, constitutes a major contribution of this article.

Following usual notational conventions, we use circles to diagram mental spaces—the cognitive domains between which mappings and links are automatically established as we think and talk. They contain elements (represented by lower case letters), connectors (represented by lines) that relate elements across spaces based on identity, analogy, representation, etc., and links (represented by straight dashed lines) between mental spaces. The starting point for any mental space configuration is the *base* space, and subordinate mental spaces are set up in the presence of *space builders*, i.e., language forms that point to conceptual domains (perspectives) like time, beliefs, wishes, movies, or pictures. Counterparts of elements created in other spaces are represented by the same letter with a subscript number. The Access Principle defines a general procedure for accessing elements: "If two elements a and a_1 are linked by a connector $F(a_1 = F(a))$, then element a_1 can be identified by naming, describing, or pointing to, its counterpart a."

Example (4), shown in Fig. 8.2, is borrowed from FAUCONNIER (1985) and provides a succinct explanation of how mental space configurations are built up.

(4) In the movie Orson Welles played Hitchcock, who played a man at the bus stop.

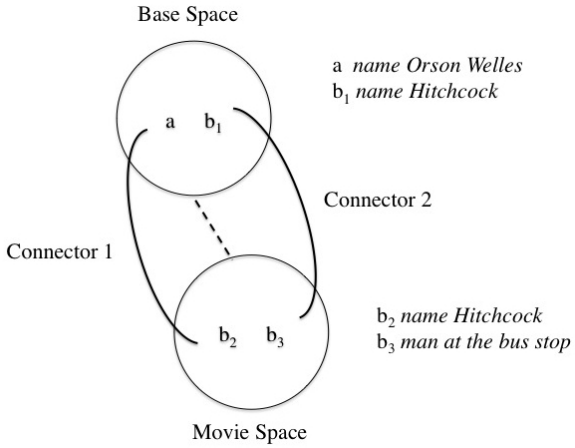

Figure 8.2: Mental space configuration of (4)

The base is always placed at the top and linked to its child spaces by a subordination relation. In this case, the base represents the reality space with the two DEs introduced by *Orson Welles* and *Hitchcock*. In addition, the two characters played by these two actors appear in the movie space, giving rise to two additional DEs. Then, Orson Welles-the-person is linked with Hitchcock-the-character (Connector 1), and Hitchcock-the-person is linked with the man at the bus stop (Connector 2). The two connectors exemplify actor-character relations.

Note that we could add a third connector linking Hitchcock-the-person (b_1) with Hitchcock-the-character (b_2), as this is a link—of the representation type—that we would make for a coherent discourse. With such a framework then, the different granularity levels at which DEs can be conceived can be easily represented by adding subordinate mental spaces with counterparts to DEs in a previous space (i.e., DEs constructed earlier in the ongoing linguistic exchange). By setting up a movie space, the discourse context in (4) turns the granularity level of person versus representation into a relevant one. In the diagrams we only show the mental spaces that are activated according to the discourse interests. That is to say, the same elements placed in another discourse could give rise to a different mental space configuration.

8.3.3 Continuum

A mental space, representing a coherent perspective on some portion of the (possibly partly imaginary) world, contains the entities (and events, etc.) present in that portion. Each entity is conceptualized by discourse participants with a set of associated features with specific values characteristic to the particular space. According to the traditional definition of coreference, entities with the same feature values are coreferent (Fig. 8.3(a)) while entities with different feature values are not (Fig. 8.3(e)). There is, however, a third, in-between possibility, namely that

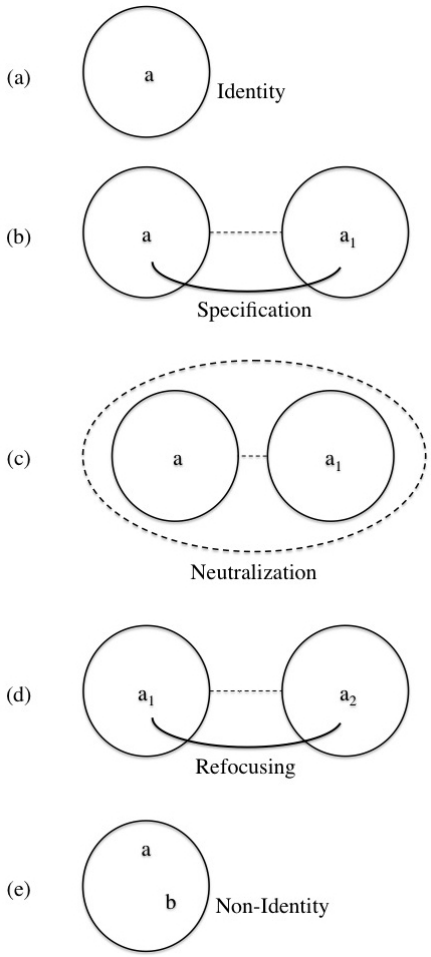

Figure 8.3: Mental space configurations of the coreference continuum

entities share most but not all feature values, and this is our main concern in this article and the reason for assuming a continuum model of coreference. One arrives to this middle-ground domain of *near-identity* by exclusion: if a relation does not fall into either identity or non-identity, then we are confronted by a near-identity relation (Fig. 8.3(b)-(d)). We claim that three different cognitive operations of categorization (specification, refocusing and neutralization) underlie near-identity relations. They are presented in Section 8.3.4.

Throughout a discourse, some DEs are mentioned multiple times and new features might be introduced, old features might be omitted or their values changed, etc. The speaker states a series of feature–value pairs that the hearer is able to recognize or know as (supposedly) true of the DE (at that time), enough to pick it out uniquely. The problem of coreference is determining whether a new expression in a discourse refers to the 'same' prior DE, or whether it introduces a new (albeit

possibly very closely related) one. Rephrased in terms of the mental space framework, the problem of coreference is determining the (in)compatibility between the feature values of the various elements and counterparts in other spaces. As we will show, feature values can be different but only potentially incompatible, where the decision is a contextual one. In our continuum model, the configuration of mental spaces is guided by two main principles:

1. Linguistic expressions (e.g., temporal phrases) that involve a change in a feature value (e.g., time) function as space builders.

2. The pragmatic context suggests a preference for feature value compatibility (or not), and hence for identity, near-identity, or non-identity of reference.

A taxonomy of the types of features that most frequently require the building of a new subordinate space when their value is changed is presented in Section 8.4. As we will show, most of the features are typical of the entities typified as Individuals by FRAURUD (1996). Being the entities that are conceived of in their own right and of which we possess token knowledge, Individuals are prone to be construed at different granularities. How two different values for the same feature compare is constrained by the pragmatic context, which can either emphasize or collapse the value difference.

Explaining coreference by co-existence of different mental spaces and their complex interplay as discourse unfolds overcomes the shortcomings of the traditional categorical definition of coreference in naturally occurring data.

8.3.4 Specification, refocusing and neutralization

Specification and neutralization are two operations of categorization that work in opposite directions. Specification, which adds features, is a shift towards greater granularity, while neutralization, which removes them, is a shift towards less granularity. The former generates from an entity one or more finer-grained entities by adding features (that are however consistent with the original), thereby creating new indexical features. Neutralization, on the other hand, blends and conflates two or more similar entities into a more general or vague category by removing features, thereby neutralizing potential indexical features. Finally, the refocusing operation, similar to specification, adds more features, but ones whose values override the original's existing (assumed) values in ways that are not consistent, thereby creating new indexical features that may or may not be more specified than the original. These three operations are best illustrated with the Postville and Jews examples from Section 8.1, repeated in (5) and (6).

(5) On homecoming night *Postville* feels like Hometown, USA, but a look around *this town of 2,000* shows it's become a miniature Ellis Island. This was an all-white, all-Christian community ... For those who prefer *the old Postville*, Mayor John Hyman has a simple answer.

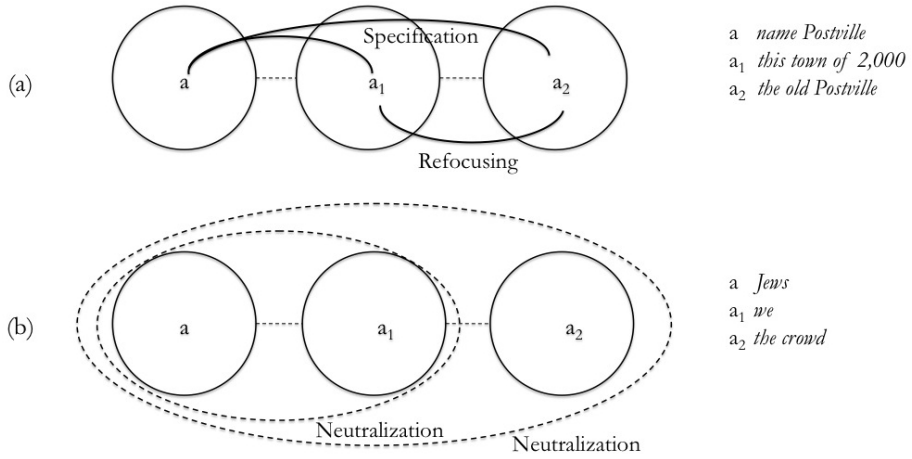

Figure 8.4: Mental space configurations of (5) and (6)

(6) Last night in Tel Aviv, *Jews* attacked a restaurant that employs Palestinians. "*We* want war," *the crowd* chanted.

In (5), one entity is *Postville*, whose name feature carries the value 'Postville' (Fig. 8.4(a)). The second mention (*this town of 2,000*) predicates a new property of an existing entity. Since mental spaces are defined as a particular (value-defined) perspective over the constituent entities, etc., it is in the nature of the theory of mental spaces that when one introduces a new value for a feature, one must, by construction, generate a new subordinate space. The citizens number feature specifies detail that is consistent with the existing DE as defined so far. This value augmentation is what we call 'specification.' The past time value of the third mention (*the old Postville*), however, clashes with the implicit time feature of the previous DE, which carries the value 'the present.' This value replacement occurs with 'refocusing.' Changing the time value from 'the present' to 'the past' for the Postville entity automatically brings into existence the new-Postville space that contains the updated Postville entity.

One aspect of entities and features makes the operation of mental spaces more complex. Some features may be underspecified or take multiple values, as occurs with the Jews example (6). The introductory entity 'Jews' is a conceptual set and hence has a members feature with values $\{\text{person}_1, \text{person}_2, \ldots, \text{person}_n\}$. The subsequent mentions *we* and *the crowd* also have a members feature, but the key issue here is not whether every member of the collection is present in all three values, but rather the set itself. For the purposes of this paper, it is irrelevant whether those who chanted are a subgroup or all of those who attacked the restaurant, or whether one of the individuals who chanted was not Jewish. Thus, we say that the three mentions have been 'neutralized' by losing a distinctive value (Fig. 8.4(b)).

These two examples serve to illustrate the role of context. When the feature value changes for communicative purposes, like in (5), where the city of Postville

is split in temporal slices to draw a distinction between the old and the new city, then we are in front of a refocusing shift between a_1, a_2, etc. (Fig. 8.4(a)). In contrast, a neutralization shift occurs when the change in value has no goal other than to present a new perspective or subsection of the old one in such a way that the feature ceases to be distinctive, and a, a_1, a_2, etc., blend together (Fig. 8.4(b)).

8.4 Types of (near-)identity relations

In this section we describe a study that identifies the (types of) features, organized into a hierarchy, that require mental space shifts when their values are changed (RECASENS *et al.*, 2010*a*). We distinguish ten different features that give rise to specification, refocusing or neutralization depending on whether the context implies a value augmentation, a value replacement, or a value loss.[6] Although this is not an exhaustive typology, it does provide the main features the change of whose values results in a near-identity relation. We arrived at this typology by a bottom-up process, first extracting problematic coreference relations[7] from real data, and then comparing inter-annotator agreement and readjusting the classes.

1. **Name metonymy**. Proper nouns naming complex Individuals are space builders when used metonymically, as they have at least one feature that can take different values depending from which facet(s) the DE is seen. For example, a company produces a product, is headquartered in a location, employs a president, etc. Under Name metonymy, a proper noun places an element in the base space, and one or more subsequent NPs refer(s) to facet(s) of the DE. Since the specific facets available depend on the type of entity under consideration, there are a great many possibilities. Nonetheless, certain metonymies occur frequently enough that we name here a few subtypes.

    1a. **ROLE**. A specific role or function performed by a human, animal or object, makes it possible to split an entity along the role feature. This can be professional (paid, as in *teacher*), non-professional (unpaid, as in *comforter*), or kinship (as in *aunt*). In (7-a), the actor (a_1) and father (a_2) pertain to two different roles of the same individual *Gassman* (a). The opposition expressed in the citation pertains to the typical activities of Gassman (actor-like actions versus father-like ones) and so causes a complete value replacement. The refocusing relation is displayed in Fig. 8.5(a). In contrast, the context presented in (7-b), which does not make the Gassman-the-actor alteration relevant but simply adds more detail, results in the mental space configuration of

[6] To avoid confusions, note that we are not listing ISA classes, but the types of the near-identity classes. These are conceptually different things.

[7] By *problematic* we mean those cases that involved disagreements between the annotators or that could be argued either way—coreferent or non-coreferent—according to the authors.

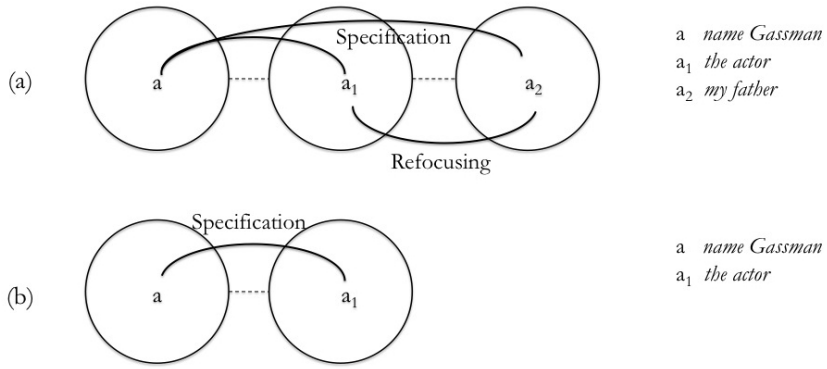

Figure 8.5: Mental space configurations of (7-a) and (7-b)

Fig. 8.5(b), where a and a_1 are only related by specification and no third mental space needs to be introduced.

(7) a. "Your father was the greatest, but he was also one of us," commented an anonymous old lady while she was shaking Alessandro's hand—[Gassman]$_a$'s best known son. "I will miss [the actor]$_{a_1}$, but I will be lacking [my father]$_{a_2}$ especially," he said.

b. Hollywood beckoned and [Gassman]$_a$ was put under contract at MGM but the studio didn't know how best to exploit [the actor]$_{a_1}$'s capabilities.

1b. **LOCATION**. As a meta-concept, the name of a location triggers a feature that can be filled with facet(s) like the physical place, the political organization, the population, the ruling government, an affiliated organization (e.g., a sport team), an event celebrated at that location, or a product manufactured at that location. In (8), the first mention of Russia (a) can be a metonymic for the political organization, the government, etc., whereas a_1 presents a more specified mention that explicitly refers to the government (Fig. 8.5(b)-like).

(8) Yugoslav opposition leaders sharply criticized both the United States and [Russia]$_a$ today as a general strike against President Slobodan Milosevic gained momentum ... Kostunica accused [the Russian government]$_{a_1}$ of indecision.

1c. **ORGANIZATION**. As a meta-concept, the name of a company or other social organization triggers a feature that can be filled with facet(s) like the legal organization itself, the facility that houses it, its shares on the stock market, its people or employees, a product that it manufactures, or an associated event like a scandal. Note that near-identity is what

licenses in (9) the use of a pronoun referring to the drink despite the fact that its antecedent refers to the company. The unreconcilable features of a_2 result in a refocusing relation (Fig. 8.5(a)-like).

(9) [Coca-Cola]$_{a_1}$ went out of business, which John thought was a shame, as he really enjoyed drinking [it]$_{a_2}$.

1d. **INFORMATIONAL REALIZATION**. An Individual corresponding to an informational object (e.g., story, law, review, etc.) contains a format feature that specifies the format in which the information is presented or manifested (e.g., book, movie, speech, etc.). In (10), refocusing explains the near-identity relation between the movie (a_1) and the book (a_2), which clash in their format value but are identical in their content, the story (Fig. 8.5(a)-like).

(10) She hasn't seen [Gone with the Wind]$_{a_1}$, but she's read [it]$_{a_2}$.

2. **REPRESENTATION**. Representational objects (pictures, statues, toy replicas, characters, maps, etc.) have a real/image feature as they generate, for an entity X, two mental spaces containing respectively Real-X and Image-X. For Image-X to be a representation of Real-X, JACKENDOFF (1983, pg. 221) points out two preference rules: (i) dubbing, by which the creator of the image has stipulated the entity in question as an Image-X, and (ii) resemblance, by which Image-X must somehow look like Real-X. There can be more than one Image-X, like in (11), where a_2 replaces the image value of a_1 (Fig. 8.5(a)-like). The representation can also be of a more abstract kind, like one's mental conceptualization of an object.

(11) We stand staring at two paintings of [Queen Elizabeth]$_a$. In the one on the left, [she]$_{a_1}$ is dressed as Empress of India. In the one on the right, [she]$_{a_2}$ is dressed in an elegant blue gown.

3. **Meronymy**. The different value of the constitution feature (e.g., parts, composition, members) between meronyms and holonyms can be neutralized in a near-identity relation. Inspired by CHAFFIN *et al.* (1988), we distinguish the following three main subtypes.

3a. **PART·WHOLE**. It is possible for an entity whose parts feature value carries a functionally relevant part of another entity to neutralize with the latter. In (12), President Clinton (a) is seen as a functioning part of the entire US government (a_1). By neutralizing them we drop those features of Clinton that make him a person and keep only those that make him a government functionary (Fig. 8.6).

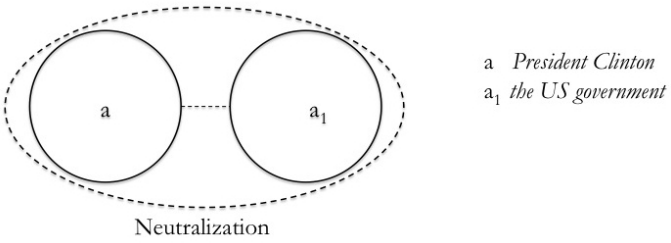

Figure 8.6: Mental space configuration of (12)

(12) Bangladesh Prime Minister Hasina and [President Clinton]$_{a}$ expressed the hope that this trend will continue ... Both [the US government]$_{a_1}$ and American businesses welcomed the willingness of Bangladesh to embrace innovative approaches towards sustainable economic growth.

3b. **STUFF·OBJECT**. It is possible for a DE to neutralize with another DE if the composition feature value of one carries the main constituent material of the other. Unlike components, the stuff of which a thing is made cannot be separated from the object. Given that the most relevant component of alcoholic drinks is alcohol, the two can be neutralized, as in (13), to refer to the 'same' (Fig. 8.6-like).

(13) The City Council approved legislation prohibiting selling [alcoholic drinks]$_{a}$ during night hours ... Bars not officially categorized as bars will not be allowed to sell [alcohol]$_{a_1}$.

3c. **OVERLAP**. When two DEs denote two overlapping (possibly unbounded) sets, discourse participants intuitively neutralize the members feature as near-identical even though they might not correspond to exactly the same collection of individuals. Unlike PART·WHOLE (3a), the collection consists of repeated, closely similar, members, and the members are not required to perform a particular function distinct from one another. The Jews example above (6) as well as (14) belong here (Fig. 8.6-like).

(14) [An International team]$_{a}$ is developing a vaccine against Alzheimer's disease and [they]$_{a_1}$ are trying it out on a new and improved mouse model of the onus.

4. **Spatio-temporal function**. Temporal and locative phrases change the space or time feature of an entity: it is the 'same' entity or event but realized in another location or time. Accordingly, we differentiate the following two subtypes.

4a. **PLACE**. If a DE is instantiated in a particular physical location, it generates a more specified DE with a specific place feature value. It is possible for the fine-grained copies to coexist but not in the same place. Although the two NPs in (15) refer to the same celebration, the first place feature carries the value 'New York' while the second refocuses the value to 'Southern Hemisphere' (Fig. 8.5(a)-like).

(15) [New York's New Year's Eve]$_{a_1}$ is one of the most widely attended parties in the world ... Celebrating [it]$_{a_2}$ in the Southern Hemisphere is always memorable, especially for those of us in the Northern Hemisphere.

4b. **TIME**. Different subordinate DEs are created due to a change in the time feature value, which is underspecified in the base DE. Seeing an object as a set of temporal slices, each subordinate DE represents a slice of the object's history, like the Postville example (5). It is not possible for the temporally-different DEs to coexist. Note that *the night* in (16) can be replaced with *this year's New Year's Eve* (Fig. 8.5(a)-like).

(16) After the extravagance of [last year's New Year's Eve]$_{a_1}$, many restaurants are toning things down this year, opting for a la carte menus and reservations throughout [the night]$_{a_2}$.

Spatio-temporal near-identity typically results from a numerical function (17-a) or a role function (17-b). Subordinate DEs refer to either the same function (e.g., price, age, rate, etc.) or the same role (e.g., president, director, etc.) as the base DE, but have different numerical values or are filled by a different person due to a change in time, space or both.

(17) a. At 8, [the temperature]$_{a_1}$ rose to 99º. This morning [it]$_{a_2}$ was 85º.
b. In France, [the president]$_{a_1}$ is elected for a term of seven years, while in the United States [he]$_{a_2}$ is elected for a term of four years.

8.5 Stability study

As part of the bottom-up process of establishing the most frequent types of near-identity relations (Section 8.4), we carried out three annotation experiments on a sample of naturally occurring data. They helped identify weaknesses in the typology and secure stability of the theoretical model. The last experiment established inter-annotator agreement at acceptable levels: $\kappa = 0.58$ overall, and up to $\kappa = 0.65$ and $\kappa = 0.84$ leaving out one and two outliers, respectively. We briefly summarize these previous experiments and discuss the results, as they led us to the idea that

different values for the same feature do not only relate in a near-identity way but also in an either specification, refocusing or neutralization direction. Most of the remaining disagreements were explainable in these terms. The study as a whole provided evidence that coreference is best approached as a continuum.

8.5.1 Method

8.5.1.1 Participants

Six paid subjects participated in the experiments: four undergraduate students and two authors of this paper. Although the undergraduates were not linguistics students, they were familiar with annotation tasks requiring semantic awareness, but had not worked on coreference before.

8.5.1.2 Materials

A total of 60 text excerpts were selected from three electronic corpora—ACE (DODDINGTON *et al.*, 2004), OntoNotes (PRADHAN *et al.*, 2007*a*) and AnCora (RECASENS i MARTÍ, 2010)—as well as from the Web, a television show, and real conversation. The excerpts were divided in three groups of 20, each including examples of the different coreference types in different proportions so that annotators could not reason by elimination or the like. To the same effect, each round varied the proportions, with a mean of 27% identity, 67% near-identity, and 6% non-identity cases. The largest number of examples always was near-identity because this was our main interest. In each excerpt, two or more NPs were marked with square brackets and were given a subscript ID number. Apart from the set of 20 excerpts, annotators were given an answer sheet where all the possible combinatorial pairs between the marked NPs were listed. The first 20 excerpts included 78 pairs to be analyzed; the second, 53, and the third, 43. The excerpts that were used are collected in Appendix B.[8]

8.5.1.3 Procedure

The task required coders to read the annotation guidelines and classify the selected pairs of NPs in each excerpt according to the (near-)identity relation(s) that obtained between them by filling in the answer sheet. They had to assign one or more, but at least one, class to each pair of NPs, indicating the corresponding type and subtype identifiers. They were asked to specify all the possible (sub)types for underspecified pronouns and genuinely ambiguous NPs that accepted multiple interpretations, and to make a note of comments, doubts or remarks they had. The three groups of 20 excerpts were annotated in three separate experiments, spread out over a span of four weeks. In each experiment, a different version of the anno-

[8]We include the entire collection of selected texts in the appendices as they make evident the limitations of a categorical definition of coreference as well as the difficulty of the task.

tation guidelines was used, since the typology underwent substantial revision—in a decreasing manner—after completing each round.

8.5.2 Results and discussion

Inter-coder agreement was measured with Fleiss's kappa (FLEISS, 1981), as it can assess the agreement between more than two raters, unlike other kappas such as Cohen's kappa. The measure calculates the degree of agreement in classification over that which would be expected by chance and its values range between -1 and 1, where 1 signifies perfect agreement, 0 signifies no difference from chance agreement, and negative values signify that agreement is weaker than expected by chance. Typically, a kappa value of at least 0.60 is required. For the cases in which a coder gave multiple relations as an answer, the one showing the highest agreement was used for computing kappa. Kappa was computed with the R package *irr*, version 0.82 (GAMER *et al.*, 2009). Statistical significance was tested with a kappa z-test provided by the same package.

8.5.2.1 Experiment 1

The 20 texts used in this first experiment, which served as a practice round, are included in Appendix B.1. After counting the number of times a type was assigned to each pair of NPs, we only obtained overall $\kappa = 0.32$. More importantly, this first experiment revealed serious shortcomings of the first version of the typology. In this regard, the comments and notes included by the coders in the answer sheet were very helpful.

Very few cases obtained high agreement. We were surprised by pairs such as (4, 2-3) and (15, 1-2)[9] for which coders selected four—or even five—different types. At this early stage, we addressed most of the disagreements by including additional types, removing broad ones without identifying force, or restricting the scope of existing ones. We also improved the definitions in the guidelines, as they generally lacked criteria for choosing between the different types.

Interestingly, we observed that most relations with two-type answers included a near-identity type and either IDENTITY (6, 1-3) or NON-IDENTITY (6, 2-3). Apart from supporting the continuum view, this was later the inspiration to distinguishing two main directions within near-identity: relations perceived on the borderline with identity would fall into either specification or neutralization, whereas those perceived on the borderline with non-identity would fall into refocusing. On the other hand, some relations with varied answers were indicative of the multiplicity of interpretation—and thus the difficulty of a categorical classification task. It is in this same regard that we interpreted multiple answers given by the same annotator, the highest number of types being three. In (14), the types given to the NP pairs

[9]References to the excerpts in Appendix B are as follows: first the excerpt number, and then the ID numbers of the two NPs whose relation is under analysis.

got swapped between coders: coder *a* interpreted (14, 1-2) as an IDENTITY relation and coder *b* as a TIME relation, but vice versa for (14, 1-3). It was mostly an effect of the underspecified nature of the pronoun. Note that this disagreement can be better accounted for under the light of neutralization.

8.5.2.2 Experiment 2

As a result of revising the guidelines after Experiment 1, the agreement of the second set of 20 texts (Appendix B.2) reached $\kappa = 0.54$. In contrast with Experiment 1, the answers were not so spread over different types. To address the low disagreement obtained by a few types, a solution was found in setting clear preferences in the guidelines for cases when it was possible for two near-identity classes to co-occur, as more than one feature value can change and still be perceived as near-identity, e.g., LOCATION and PART·WHOLE (29, 1-2).

Again, some of the pairs with two- or three-type answers manifested different mental space configurations compatible for the same discourse, as some cases accepted more than a single viewpoint, e.g., ROLE and REPRESENTATION (27, 1-2). Similarly, some of the isolated (5-to-1) answers revealed yet another—though less frequent—interpretation (40, 1-2). A large number of isolated answers, however, made us consider the possible presence of outliers, and we detected two. If agreement was computed between the other five coders, we obtained a considerable improvement resulting in $\kappa = 0.63$; between the other four coders, $\kappa = 0.71$.

8.5.2.3 Experiment 3

The final set of 20 texts (Appendix B.3) obtained a further improvement in agreement, $\kappa = 0.58$, as shown by the kappa scores in Table 8.1, and up to $\kappa = 0.65$ and $\kappa = 0.84$ leaving out the one and two outliers, respectively. The changes introduced in the typology after Experiment 2 were small compared with the revision we undertook after Experiment 1. Basically, we improved the guidelines by adding some clarifications and commenting all the examples. Nevertheless, the (near-)identity task is difficult and requires a mature sensitivity to language that not all coders had, as revealed by the presence of outliers.

The fact that this third experiment showed a lower number of many-type-answer relations, an insignificant number of relations with both IDENTITY and NON-IDENTITY answers, but still a high number of two-type-answer relations, most of them including a near-identity type and either IDENTITY or NON-IDENTITY, led us to conclude that disagreements were mainly due to the fuzziness between borderline identity types rather than to the typology of near-identity types. It emerged that the feature values were not always interpreted uniformly by all coders: near-identical for some, and simply identical or non-identical for others. At this point we took the decision of dividing the middle ground of the coreference continuum into three directions—specification, neutralization and refocusing—in order to have three umbrella terms for such borderline cases.

Relation	Type	Subtype	Kappa	z	p-value
1. Non-Identity			0.89	22.64	0.00
2. Identity			0.30	7.55	0.00
3. Near-Identity	A. Name metonymy	a. Role	-0.00	-0.10	0.92
		b. Location	0.87	22.01	0.00
		c. Organization	0.48	12.09	0.00
		d. Information realization	0.49	12.54	0.00
		e. Representation	0.59	15.08	0.00
		f. Other	0.59	15.08	0.00
	B. Incidental meronymy	a. Part·Whole	-0.00	-0.10	0.92
		b. Stuff·Object	0.80	20.22	0.00
		c. Overlap	0.73	18.44	0.00
	C. Class	a. More specific	0.39	9.80	0.00
		b. More general	0.38	9.61	0.00
	D. Spatio-temporal function	a. Place	0.67	16.90	0.00
		b. Time	0.70	17.70	0.00
		c. Numerical function			
		d. Role function	-0.01	-0.20	0.84
Total			0.58	39.50	0.00

Table 8.1: Results of Experiment 3

The limitations of categorical approaches were manifested again by cases accepting multiple interpretations, which is in accordance with the predictions of mental space theory. One feature type tends to prevail, as shown by the large number of isolated answers, but there were a few 50%–50% cases. For instance, (59, 1-2) included four OVERLAP answers, four TIME, and one NON-IDENTITY. It revealed the fact that discourse participants do not always conceptualize entities in the same way: while *the people* and *they* could be two groups with overlapping members, they could also have two different time features (the people from the past versus the people from today).

The general conclusion we drew was that regarding coreference in terms of a continuum is certainly the most explanatory approach: there are prototypical examples illustrating clearly each identity type but also a wide range of intermediate cases accepting interpretations from varied angles, depending on the dimension that is felt as dominant. The typology presented in Section 8.4 is a compact version that does away with the too specific, redundant, types of Table 8.1.

8.6 Conclusion

We discussed the shortcomings of a categorical understanding of coreference as it is too limited to take into account the role of cognitive processes in the dynamic interpretation of discourse, and hence leads to contradictory analyses and annotation. It fails when confronted with the full range of natural language phenomena.

The complexity of coreference becomes apparent once we reject the naive view of linguistic expressions as mere pointers to a unique objective world, and acknowledge that the categories and concepts of our mental apparatus rely on a projected world. Discourse constructs its own model with its own entities, which language users conceptualize at a coarser or more fine-grained granularity depending on the communicative purpose. In discourse, identity behaves in a fashion different from mathematical or logical identity. Accordingly, we argued for a continuum approach to coreference that contemplates middle-ground relations of near-identity, which make complete sense in the framework of Fauconnier's mental space theory. Near-identity appears to be key to describe those relations between elements of different spaces that share most but not all feature values.

Three inter-annotator agreement studies provided further evidence for a continuum approach to coreference and led us to distinguish the main types of features that typically result in near-identity relations when their value differs. In addition, we identified three major cognitive operations of categorization depending on whether there is an expansion of a feature value (specification shift), a complete value replacement (refocusing shift), or a loss of a distinctive value (neutralization shift). The fact that it is possible for the same relation to be explained by a change in different feature types is a direct reflection of the rich and varied categorization process that underlies discourse interpretation, thus suggesting that any effort to impose limitations to one single type is likely to fail. Rather, our framework is best viewed as a set of directions and tendencies that help interpret how coreference phenomena occur in discourse under the understanding that there are no absolute and universal rules.

Acknowledgements We are grateful to Jerry Hobbs for his valuable insights, and to the annotators: David Halpern, Peggy Ho, Justin James, and Rita Zaragoza.

This work was supported in part by the Spanish Ministry of Education through an FPU scholarship (AP2006-00994) and the TEXT-MESS 2.0 Project (TIN2009-13391-C04-04).

$\star \star \star$

CAPÍTOL 9

Conclusions i perspectives de futur

En aquest últim capítol, traço una retrospectiva dels assoliments d'aquesta tesi. L'èxit principal ha estat expandir el coneixement de la coreferència proposant una visió més àmplia del problema i posant de manifest les diverses barreres que actualment obstaculitzen el bon funcionament dels sistemes de resolució automàtica de la coreferència. Aquesta tesi ha avançat en la comprensió de qüestions clau relacionades amb l'enfocament teòric, l'anotació de corpus, el tractament computacional i l'avaluació del problema de la coreferència. He replantejat el problema en termes lleugerament diferents per adequar-lo a la realitat lingüística.

El capítol comença amb una breu descripció de les aportacions més destacades d'aquest treball i una avaluació de les lliçons apreses (apartat 9.1) i, a continuació, es presenten algunes idees interessants i reptes sorgits de la meva anàlisi per abordar en futures investigacions (apartat 9.2).

9.1 Conclusions

Aquesta tesi reuneix diverses contribucions a diferents facetes del problema de la coreferència. Desglossat per facetes, destacaria:

- Anotació
  - el desenvolupament d'una guia d'anotació de coreferència per a dades del català i el castellà.
  - la construcció del corpus AnCora-CO, el corpus més gran del català i el castellà anotat amb relacions de coreferència.[1]

[1] http://clic.ub.edu/corpus/ancora

- Resolució
  - l'establiment de més de quaranta-cinc trets d'aprenentatge per a la resolució de la coreferència en castellà i l'anàlisi de la seva contribució en un model basat en parells.
  - la presentació de CISTELL, un sistema de resolució de la coreferència comparable als sistemes de l'estat de l'art i que permet utilitzar coneixement del discurs i contextual així com trets a nivell de *cluster*.
  - l'organització i preparació dels recursos per a la primera tasca compartida del SemEval-2010 en "Resolució de la coreferència en múltiples llengües".[2]
- Avaluació
  - una anàlisi comparativa i detallada de les diferents mètriques d'avaluació per a la coreferència: MUC, B^3, CEAF, valor ACE, F1 per parells, informació mútua i índex de Rand.
  - l'elaboració de la mètrica d'avaluació BLANC per resoldre el problema de les entitats unàries que presenten les altres mesures. Aquesta mètrica adapta l'índex de Rand per tal de recompensar equitativament els enllaços de coreferència i els de no coreferència.
- Teoria
  - la delimitació de les similituds i diferències entre els conceptes de coreferència i paràfrasi.
  - la defensa de la necessitat d'introduir el concepte de *quasi-identitat* a l'anàlisi discreta habitual de la coreferència.
  - la construcció del primer corpus de textos reals amb casos de quasi-identitat. La definició d'una tipologia de relacions de quasi-identitat i un estudi d'acord entre anotadors.
  - la presentació d'un nou model de la coreferència entesa com a contínuum així com tres operacions cognitives de categorització: l'especificació, la neutralització i el reenfocament.

La vàlua de les contribucions d'aquesta tesi radica, sobretot, en examinar els supòsits subjacents a la gran majoria d'investigacions anteriors i aclarir les qüestions per resoldre. Pel que fa a l'anotació, he sostingut que ha d'estar recolzada en una **teoria més completa de la coreferència**. Necessitem una teoria que expliqui els complexos patrons trobats a les dades reals com ara les relacions que no es poden classificar clarament ni com a coreferents ni com a no coreferents i les mencions que es troben al pont entre la referencialitat i la no referencialitat. El model continu de la coreferència que he presentat ofereix un marc teòric adequat per examinar

[2] http://stel.ub.edu/semeval2010-coref/

el caràcter no discret de la coreferència i, en certa mesura, també de la referencialitat. En l'actualitat, la manca d'un **veritable corpus que sigui l'estàndard de referència** té serioses implicacions per al desenvolupament de sistemes de resolució de la coreferència i, sobretot, per a la seva comparació. He advocat pels esforços d'anotació que identifiquen el conjunt de mencions amb el conjunt complet de SNs referencials –la qual cosa exclou els sintagmes atributius i predicatius– i que anoten tant les entitats de múltiples mencions com les entitats unàries.

Una segona limitació dels mètodes existents està lligada amb les mancances del trets d'aprenentatge actualment utilitzats, que són a penes generalitzables i que no inclouen **coneixement pragmàtic i del món**. Codificar aquest tipus de coneixement d'alguna forma per tal que un sistema de resolució de la coreferència el pugui utilitzar és la clau per avançar substancialment en l'estat de l'art. Donada l'extensa casuística i diversitat d'entorns en què es poden donar les relacions de coreferència, tenir en compte el context més que no pas les mencions aïllades és també essencial. El sistema CISTELL que he desenvolupat ofereix l'oportunitat d'emmagatzemar i transportar al llarg del procés de resolució no només la informació sobre una menció que es troba "dins" del text, sinó també coneixement contextual i del món de "fora" del text. Només així podrem aconseguir que el sistema tingui la informació necessària per decidir, per a cadascuna de les mencions, si fa o no fa referència a una entitat particular. Encara que els mètodes basats en aprenentatge poden ser de gran ajut durant aquest procés, convé que siguin **dissenyats manualment i a consciència**, no aplicats a cegues.

Finalment, els principals problemes amb les pràctiques actuals d'avaluació de la coreferència provenen de **biaixos en les mètriques més comunament utilitzades** (B^3, CEAF, MUC). Per això, per exemple, els *baselines* de totes-unàries i de mateix-nucli són difícils de superar quan els sistemes s'avaluen sobre un corpus anotat amb tot el conjunt de mencions. Això s'agreuja pel fet que la major part dels sistemes no acostumen a ser avaluats qualitativament ni se'n mostren les sortides. La mesura que he proposat, BLANC, té com a objectiu buscar una **solució que trobi l'equilibri entre el gran nombre d'entitats unàries i el nombre relativament petit d'entitats de múltiples mencions**. A més, he argumentat que avaluar la identificació de mencions i la resolució de la coreferència amb una única puntuació pot resultar enganyós en determinades situacions. Com a alternativa, he suggerit utilitzar mencions reals i separar la tasca d'identificació de mencions com una tasca en si mateixa. Per últim, les xifres soles no són suficients per mesurar el funcionament d'un sistema i posar a diposició de la comunitat científica les sortides dels sistemes, com s'ha fet amb els que van participar al SemEval, hauria de ser pràctica comuna.

En última instància, s'espera que la recerca descrita en aquesta tesi contribueixi a trobar maneres més eficaces d'abordar la tasca de resolució de la coreferència. Ja he fet algun pas en aquesta direcció i tinc idees per seguir-hi treballant en futures investigacions. Aquest és el nucli del següent apartat.

9.2 Perspectives de futur

Com a resultat de les converses mantingudes amb diversos investigadors i de l'anàlisi duta a terme en el transcurs d'aquest projecte, han sorgit possibles línies de recerca per ampliar la feina que he presentat. Aquest apartat resumeix les més significatives, algunes de les quals ja s'han esmentat en algun punt de la tesi. Els reptes que planteja avui la coreferència es poden resumir en dues qüestions principals: (1) el món de les entitats discursives, la seva representació i el seu comportament i (2) la naturalesa i el funcionament d'un sistema de coreferència que sigui dinàmic i no discret.

Pel que fa al món de les entitats discursives, el model continu de la coreferència presenta molts avantatges, incloent la cobertura i la robustesa. Cal seguir treballant en determinar quins atributs i dimensions tenen un paper important, donades les necessitats de la semàntica i el context, així com en determinar els valors de cada atribut a partir del text, el context i el coneixement del món (per exemple, la web, la Viquipèdia, bases de dades, etc.). Aquesta és la informació a què el sistema CISTELL hauria d'accedir per tal d'incloure el tipus de coneixement contextual i del món que manca als sistemes de resolució de la coreferència actuals. Tal com NG (2010) assenyala en el seu estudi, els mètodes no supervisats competeixen amb els seus homòlegs supervisats, la qual cosa posa en dubte "si els sistemes supervisats estan fent un ús realment eficient de les dades etiquetades disponibles". El problema prové del fet que els textos no fan explícita tota la informació que es requereix per a la seva comprensió (però que les persones recuperem sense esforç). Una manera fiable d'obtenir aquesta informació és partir de treballs recents d'extracció de coneixement de la web (MARKERT i NISSIM, 2005; KOZAREVA i HOVY, 2010) o de lectura automàtica (PEÑAS i HOVY, 2010). Encara que l'extracció i la integració de tota la informació d'un text superen els límits de les capacitats actuals, s'haurien de tenir presents com un objectiu a llarg termini.

Des del punt de vista lingüístic, hi ha molt a discutir sobre què és exactament el *context*. És una noció que s'acostuma a parametritzar segons objectius empírics o teòrics. Per exemple, BACH (2005) explica que "el que s'anomena vagament 'context' és l'escenari de conversa entès en un sentit ampli: és el context cognitiu mutu o els punts rellevants compartits. Inclou l'estat actual de la conversa (el que s'acaba de dir, allò a què ens acabem de referir, etc.), l'entorn físic (si el conversants estan cara a cara), el coneixement mutu entre els conversants així com el coneixement comú rellevant a un nivell més ampli". Per tant, una altra línia de recerca interessant a llarg termini ens hauria de portar cap a una teoria comprensiva del context.

És possible generar més coneixement sobre el model continu de la coreferència i la idea de quasi-identitat mitjançant experiments psicolingüístics: Té la quasi-identitat algun efecte en el temps de processament? En quina mesura és el context determinant d'una o altra lectura? Hi ha una realitat psicolingüística al darrera de les operacions d'especificació, neutralització i reenfocament? Tant com a suport com a extensió de la investigació psicolingüística, desenvolupar un corpus

de gran tamany anotat amb relacions de quasi-identitat és un objectiu a curt termini que possibilitarà l'estudi de qüestions més concretes com ara la interacció entre la quasi-identitat i l'estructura temporal i la direccionalitat del discurs, o la interacció entre la quasi-identitat i el tipus ontològic de l'entitat. Un corpus d'aquestes característiques, a més, brindarà un conjunt de dades d'entrament i de test per al desenvolupament de sistemes més refinats de resolució de la coreferència, i animarà altres investigadors a treballar en aquesta mateixa direcció. La tipologia presentada al capítol 8 pot servir com a punt de partida per formular la guia d'anotació.

Integrar plenament el contínuum de quasi-identitat i el sistema CISTELL porta a l'altra gran qüestió: la construcció d'un sistema de resolució de la coreferència dinàmic i no discret. Una línia de recerca prometedora en aquest àmbit consisteix en establir els principis que han de permetre a una màquina de resolució de la coreferència prendre les seves decisions. Per això és necessari fer els diagrames d'espais mentals del capítol 8 més explícits en termes objectius i quantificables, així com formalitzar les operacions de neutralització, reenfocament, etc. Per a aquest objectiu, les estructures de trets tipificades i la unificació poden resultar útils.

Després de quinze anys d'investigació en coreferència basada en aprenentatge, s'ha fet evident que el model basat en parells de mencions és feble i que els models globals obtenen millors resultats (NG, 2010), però encara no està clar quina és la millor manera de dissenyar trets a nivell de *cluster* i combinar la seva informació. Desenvolupar un sistema que manipuli els cistells no estàticament sinó dinàmica és un punt lògic de partida. A llarg termini, hem d'arribar a un sistema capaç d'utilitzar el coneixement contextual i del món juntament amb les intencions dels interlocutors per inferir automàticament quins atributs són potencialment rellevants i, al seu torn, quins valors cal propagar a nous espais mentals i quins no. Per tractar aquesta qüestió, s'haurien de representar els cistells no com un conjunt de trets fixos amb valors dins o fora sinó com un conjunt de trets amb valors probabilístics de pertinença que permeti encaixos molt més fins.

La manera i l'ordre en què s'haurien d'explotar el text, el context i el coneixement del món per construir cistells estan estretament relacionats amb la manera en què s'haurien de comparar els continguts codificats en diferents cistells. Idealment doncs, la recerca sobre l'elecció i definició dels valors dels trets de cada cistell i la recerca sobre l'encaix entre cistells ha de ser en col·laboració i, en la mesura que sigui possible, coordinada.

Per últim, pel que fa a l'avaluació, l'ús de la mesura BLANC en futurs estudis per presentar els resultats de coreferència portarà, amb el temps, a millores addicionals com l'ajustament del paràmetre alfa. He mostrat els punts forts i febles de les mesures actuals, però cal impulsar la recerca de qüestions relacionades amb la definició de les fórmules de totes les mesures, com la qüestió de si són necessàries les correccions de l'atzar (VINH *et al.*, 2009) o de com es mantenen les variacions típiques dels resultats en diferents condicions i amb diferents tamanys de dades. La investigació en aquest camp es beneficiarà si la comunitat que treballa en core-

ferència adopta una mètrica estàndard d'avaluació en el futur immediat.

Després d'haver donat tant respostes com preguntes, el treball presentat en aquesta tesi acaba aquí. He respost una sèrie de preguntes, però també n'he plantejades d'altres que molt probablement seran font de motivació per a noves iniciatives d'estudi en l'apassionant camp de la coreferència.

Bibliografia

ABAD, Azad; *et al.* (2010): «A resource for investigating the impact of anaphora and coreference on inference». Dins *Proceedings of LREC 2010*, ps. 128–135. Valletta, Malta.

ALSHAWI, Hiyan (1990): «Resolving quasi logical forms». *Computational Linguistics*, volum 16(3): ps. 133–144.

AMIGÓ, Enrique; *et al.* (2009): «A comparison of extrinsic clustering evaluation metrics based on formal constraints». *Information Retrieval*, volum 12(4): ps. 461–486.

ANDROUTSOPOULOS, Ion; MALAKASIOTIS, Prodromos (2010): «A survey of paraphrasing and textual entailment methods». *Journal of Artificial Intelligence Research*, volum 38: ps. 135–187.

AONE, Chinatsu; BENNETT, Scott W. (1995): «Evaluating automated and manual acquisition of anaphora resolution strategies». Dins *Proceedings of ACL 1995*, ps. 122–129.

APPELT, Douglas; *et al.* (1995): «SRI International FASTUS System MUC-6 Results and Analysis». Dins *Proceedings of MUC-6*. Columbia, Maryland.

ARIEL, Mira (1988): «Referring and accessibility». *Journal of Linguistics*, volum 24(1): ps. 65–87.

— (2001): «Accessibility Theory: An overview». Dins *Text Representation* (Ted SANDERS; Joost SCHLIPEROORD; Wilbert SPOOREN, eds.), ps. 29–87. John Benjamins, Amsterdam.

ARNOLD, Jennifer E.; GRIFFIN, Zenzi M. (2007): «The effect of additional characters on choice of referring expression: Everyone counts». *Journal of Memory and Language*, volum 56: ps. 521–536.

ARNOLD, Jennifer E.; *et al.* (2000): «The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking». *Cognition*, volum 76: ps. B13–B26.

ARTSTEIN, Ron; POESIO, Massimo (2005): «Bias decreases in proportion to the number of annotators». Dins *Proceedings of FG-MoL 2005*, ps. 141–150. Edinburgh, Regne Unit.

— (2008): «Inter-coder agreement for computational linguistics». *Computational Linguistics*, volum 34(4): ps. 555–596.

ATTARDI, Giuseppe; ROSSI, Stefano Dei; SIMI, Maria (2010): «TANL-1: Coreference resolution by parse analysis and similarity clustering». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 108–111. Uppsala, Suècia.

AZZAM, Saliha; HUMPHREYS, Kevin; GAIZAUSKAS, Robert (1999): «Using coreference chains for text summarization». Dins *Proceedings of the ACL Workshop on Coreference and its Applications*, ps. 77–84. Baltimore, Maryland.

BACH, Kent (2005): «Context ex machina». Dins *Semantics versus Pragmatics* (Zolán Gendler SZABÓ, ed.), ps. 15–44. Clarendon, Oxford.

BAGGA, Amit; BALDWIN, Breck (1998): «Algorithms for scoring coreference chains». Dins *Proceedings of the LREC Workshop on Linguistic Coreference*, ps. 563–566. Granada, Espanya.

BAKER, Mark C. (2003): *Lexical Categories*. Cambridge University Press, Cambridge.

BALDWIN, Breck (1997): «CogNIAC: High precision coreference with limited knowledge and linguistic resources». Dins *Proceedings of the ACL-EACL Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts*, ps. 38–45. Madrid.

BARBU, Catalina; EVANS, Richard; MITKOV, Ruslan (2002): «A corpus based analysis of morphological disagreement in anaphora resolution». Dins *Proceedings of LREC 2002*, ps. 1995–1999. Las Palmas de Gran Canaria, Espanya.

BARKER, Chris (2010): «Nominals don't provide criteria of identity». Dins *The Semantics of Nominalizations across Languages and Frameworks* (Monika RATHERT; Artemis ALEXIADOU, eds.), ps. 9–24. Mouton de Gruyter, Berlín.

BEAN, David L.; RILOFF, Ellen (1999): «Corpus-based identification of non-anaphoric noun phrases». Dins *Proceedings of ACL 1999*, ps. 373–380. College Park, Maryland.

BENGTSON, Eric; ROTH, Dan (2008): «Understanding the value of features for coreference resolution». Dins *Proceedings of EMNLP 2008*, ps. 294–303. Honolulu, Hawaii.

BERGER, Adam; PIETRA, Stephen Della; PIETRA, Vincent Della (1996): «A maximum entropy approach to natural language processing». *Computational Linguistics*, volum 22(1): ps. 39–71.

BERGSMA, Shane; LIN, Dekang; GOEBEL, Randy (2008): «Distributional identification of non-referential pronouns». Dins *Proceedings of ACL-HLT 2008*, ps. 10–18. Columbus, Ohio.

BERTRAN, Manuel; *et al.* (2008): «AnCoraPipe: A tool for multilevel annotation». *Procesamiento del Lenguaje Natural*, volum 41: ps. 291–292.

BHAGAT, Rahul (2009): *Learning Paraphrases from Text*. Tesi Doctoral, University of Southern Califòrnia, Los Angeles, Califòrnia.

BLACKWELL, Sarah (2003): *Implicatures in Discourse: The Case of Spanish NP Anaphora*. John Benjamins, Amsterdam.

BORREGA, Oriol; TAULÉ, Mariona; MARTÍ, M. Antònia (2007): «What do we mean when we talk about named entities?» Dins *Proceedings of the 4th Corpus Linguistics Conference*. Birmingham, Regne Unit.

BOSQUE, Ignacio; DEMONTE, Violeta (eds.) (1999): *Gramática descriptiva de la lengua española*. Real Academia Española / Espasa Calpe, Madrid.

BOYD, Adriane; GEGG-HARRISON, Whitney; BYRON, Donna (2005): «Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated features». Dins *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, ps. 40–47. Ann Arbor, Michigan.

BRANTS, Thorsten (2000): «TnT – A statistical part-of-speech tagger». Dins *Proceedings of ANLP 2000*. Seattle, Washington.

BROSCHEIT, Samuel; *et al.* (2010): «BART: A multilingual anaphora resolution system». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 104–107. Uppsala, Suècia.

BYBEE, Joan (2010): *Language, Usage and Cognition*. Cambridge University Press, Nova York.

BYRON, Donna K. (2001): «The uncommon denominator: A proposal for consistent reporting of pronoun resolution results». *Computational Linguistics*, volum 27(4): ps. 569–578.

BYRON, Donna K.; GEGG-HARRISON, Whitney (2004): «Eliminating non-referring noun phrases from coreference resolution». Dins *Proceedings of DAARC 2004*, ps. 21–26. Azores, Portugal.

CAI, Jie; STRUBE, Michael (2010): «Evaluation metrics for end-to-end coreference resolution systems». Dins *Proceedings of SIGDIAL*, ps. 28–36. Universitat de Tòquio, Japó.

CALHOUN, Sasha; *et al.* (2010): «The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue». *Language Resources and Evaluation*, volum DOI:10.1007/s10579-010-9120-1.

CALLISON-BURCH, Chris (2007): *Paraphrasing and Translation*. Tesi Doctoral, University of Edinburgh, Edinburgh, Regne Unit.

CARBONELL, Jaime; BROWN, Ralf G. (1988): «Anaphora resolution: a multi-strategy approach». Dins *Proceedings of COLING 1988*, ps. 96–101. Budapest.

CARDIE, Claire; WAGSTAFF, Kiri (1999): «Noun phrase coreference as clustering». Dins *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, ps. 82–89. College Park, Maryland.

CARLETTA, Jean (1996): «Assessing agreement on classification tasks: the kappa statistic». *Computational Linguistics*, volum 22(2): ps. 249–254.

CARREIRAS, Manuel; GERNSBACHER, Morton Ann (1992): «Comprehending conceptual anaphors in Spanish». *Language and Cognitive Processes*, volum 7(3&4): ps. 281–299.

CHAFFIN, Roger; HERRMANN, Douglas J.; WINSTON, Morton (1988): «An empirical taxonomy of part-whole relations: Effects of part-whole relation type on relation identification». *Language and Cognitive Processes*, volum 3(1): ps. 17–48.

CHAMBERS, Nathanael; JURAFSKY, Dan (2008): «Unsupervised learning of narrative event chains». Dins *Proceedings of ACL 2008*, ps. 789–797. Columbus, Ohio.

CHAROLLES, Michel; SCHNEDECKER, Catherine (1993): «Coréférence et identité: le problème des référents évolutifs». *Langages*, volum 112: ps. 106–126.

CHOI, Yejin; CARDIE, Claire (2007): «Structured local training and biased potential functions for conditional random fields with application to coreference resolution». Dins *Proceedings of HLT-NAACL 2007*, ps. 65–72. Rochester, Nova York.

CLARK, Herbert H. (1977): «Bridging». Dins *Thinking: Readings in Cognitive Science* (P.N. JOHNSON-LAIRD; P.C.WASON, eds.), ps. 411–420. Cambridge University Press, Cambridge.

CONNOLLY, Dennis; BURGER, John D.; DAY, David S. (1994): «A machine learning approach to anaphoric reference». Dins *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP-1)*, ps. 255–261. Manchester, Regne Unit.

CRAWLEY, Rosalind; STEVENSON, Rosemary; KLEINMAN, David (1990): «The use of heuristic strategies in the interpretation of pronouns». *Journal of Psycholinguistic Research*, volum 4: ps. 245–264.

CULOTTA, Aron; *et al.* (2007): «First-order probabilistic models for coreference resolution». Dins *Proceedings of HLT-NAACL 2007*, ps. 81–88. Rochester, Nova York.

DAELEMANS, Walter; BOSCH, Antal Van den (2005): *Memory-Based Language Processing*. Cambridge University Press, Cambridge.

DAELEMANS, Walter; BUCHHOLZ, Sabine; VEENSTRA, Jorn (1999): «Memory-based shallow parsing». Dins *Proceedings of CoNLL 1999*. Bergen, Noruega.

DAUMÉ III, Hal; MARCU, Daniel (2005): «A large-scale exploration of effective global features for a joint entity detection and tracking model». Dins *Proceedings of HLT-EMNLP 2005*, ps. 97–104. Vancouver, Canadà.

DAVIES, Sarah; *et al.* (1998): «Annotating coreference in dialogues: Proposal for a scheme for MATE». `http://www.ims.uni-stuttgart.de/projekte/mate/mdag`.

DENIS, Pascal (2007): *New learning models for robust reference resolution*. Tesi Doctoral, University of Texas at Austin, Austin, Texas.

DENIS, Pascal; BALDRIDGE, Jason (2007): «Joint determination of anaphoricity and coreference resolution using integer programming». Dins *Proceedings of NAACL-HLT 2007*. Rochester, Nova York.

— (2008): «Specialized models and ranking for coreference resolution». Dins *Proceedings of EMNLP 2008*, ps. 660–669. Honolulu, Hawaii.

— (2009): «Global joint models for coreference resolution and named entity classification». *Procesamiento del Lenguaje Natural*, volum 42: ps. 87–96.

DODDINGTON, George; *et al.* (2004): «The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation». Dins *Proceedings of LREC 2004*, ps. 837–840. Lisboa.

DOLAN, Bill; BROCKETT, Chris; QUIRK, Chris (2005): «README file included in the Microsoft Research Paraphrase Corpus». Redmond, Washington.

DRAS, Mark (1999): *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Tesi Doctoral, Macquarie University, Sydney, Austràlia.

ECKERT, Miriam; STRUBE, Michael (2000): «Dialogue acts, synchronising units and anaphora resolution». *Journal of Semantics*, volum 17(1): ps. 51–89.

ELSNER, Micha; CHARNIAK, Eugene (2010): «The same-head heuristic for coreference». Dins *Proceedings of ACL 2010 Short Papers*, ps. 33–37. Uppsala, Suècia.

ERK, Katrin; STRAPPARAVA, Carlo (eds.) (2010): *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*. Uppsala, Suècia.

EVANS, Richard (2000): «A comparison of rule-based and machine learning methods for identifying non-nominal *it*». Dins *Proceedings of NLP 2000*, *LNAI*, volum 1835/2000, ps. 233–241. Springer-Verlag, Berlín.

FAUCONNIER, Gilles (1985): *Mental Spaces: Aspects of Meaning Construction in Natural Language*. MIT Press, Cambridge.

— (1997): *Mappings in Thought and Language*. Cambridge University Press, Cambridge.

FELLBAUM, Christiane (1998): *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

FERRÁNDEZ, Antonio; PALOMAR, Manuel; MORENO, Lidia (1999): «An empirical approach to Spanish anaphora resolution». *Machine Translation*, volum 14: ps. 191–216.

FINKEL, Jenny Rose; MANNING, Christopher D. (2008): «Enforcing transitivity in coreference resolution». Dins *Proceedings of ACL-HLT 2008*, ps. 45–48. Columbus, Ohio.

FLEISS, J.L. (1981): *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Nova York, segona edició.

FRAMPTON, Matthew; *et al.* (2009): «Who is "you"? Combining linguistic and gaze features to resolve second-person references in dialogue». Dins *Proceedings of EACL 2009*, ps. 273–281. Atenes.

FRAURUD, Kari (1990): «Definiteness and the processing of NPs in natural discourse». *Journal of Semantics*, volum 7: ps. 395–433.

— (1992): *Processing Noun Phrases in Natural Discourse*. Tesi Doctoral, Universitat d'Estocolm, Estocolm.

— (1996): «Cognitive ontology and NP form». Dins *Reference and Referent Accessibility* (Thorstein FRETHEIM; Jeanette K. GUNDEL, eds.), ps. 65–87. John Benjamins, Amsterdam.

FREGE, Gottlob (1892): «On sense and reference». Dins *Translations from the Philosophical Writings of Gottlob Frege* (Peter GEACH; Max BLACK, eds.), ps. 56–78. Basil Blackwell (1952), Oxford.

FUCHS, Catherine (1994): *Paraphrase et énonciation. Modélisation de la paraphrase langagière*. Ophrys, París.

FUJITA, Atsushi (2005): *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Tesi Doctoral, Institut Nara de Ciència i Tecnologia, Ikoma, Nara, Japó.

GAIZAUSKAS, Robert; *et al.* (1995): «Description of the LaSIE system as used for MUC-6». Dins *Proceedings of MUC-6*, ps. 207–220.

GAMER, Matthias; LEMON, Jim; FELLOWS, Ian (2009): *irr: Various Coefficients of Interrater Reliability and Agreement*. URL `http://CRAN.R-project.org/package=irr`, r package version 0.82.

GARIGLIANO, Roberto; URBANOWICZ, Agnieszka; NETTLETON, David J. (1997): «University of Durham: Description of the LOLITA system as used in MUC-7». Dins *Proceedings of MUC-7*.

GEACH, Peter (1962): *Reference and Generality*. Cornell University Press, Ithaca.

— (1967): «Identity». *Review of Metaphysics*, volum 21: ps. 3–12.

GERBER, Matthew; CHAI, Joyce Y. (2010): «Beyond NomBank: A study of implicit arguments for nominal predicates». Dins *Proceedings of ACL 2010*, ps. 1583–1592.

GORDON, Peter C.; GROSZ, Barbara J.; GILLIOM, Laura A. (1993): «Pronouns, names, and the centering of attention in discourse». *Cognitive Science*, volum 17: ps. 311–347.

GRISHMAN, Ralph; SUNDHEIM, Beth (1996): «Message Understanding Conference-6: a brief history». Dins *Proceedings of COLING 1996*, ps. 466–471. Copenhaguen.

GROSZ, Barbara J.; JOSHI, Aravind K.; WEINSTEIN, Scott (1995): «Centering: A framework for modeling the local coherence of discourse». *Computational Linguistics*, volum 21(2): ps. 202–225.

GROSZ, Barbara J.; SIDNER, Candace L. (1986): «Attention, intention, and the structure of discourse». *Computational Linguistics*, volum 12(3): ps. 175–204.

GUNDEL, Jeanette; HEDBERG, Nancy; ZACHARSKI, Ron (1993): «Cognitive status and the form of referring expressions in discourse». *Language*, volum 69(2): ps. 274–307.

GUODONG, Zhou; FANG, Kong (2009): «Global learning of noun phrase anaphoricity in coreference resolution via label propagation». Dins *Proceedings of EMNLP 2009*, ps. 978–986. Suntec, Singapur.

HAGHIGHI, Aria; KLEIN, Dan (2007): «Unsupervised coreference resolution in a nonparametric Bayesian model». Dins *Proceedings of ACL 2007*, ps. 848–855. Praga.

— (2009): «Simple coreference resolution with rich syntactic and semantic features». Dins *Proceedings of EMNLP 2009*, ps. 1152–1161. Suntec, Singapur.

— (2010): «Coreference resolution in a modular, entity-centered model». Dins *Proceedings of HLT-NAACL 2010*, ps. 385–393. Los Angeles, Califòrnia.

HALL, Johan; NIVRE, Joakim (2008): «A dependency-driven parser for German dependency and constituency representations». Dins *Proceedings of the ACL Workshop on Parsing German (PaGe 2008)*, ps. 47–54. Columbus, Ohio.

HALL, Johan; *et al.* (2007): «Single malt or blended? A study in multilingual parser optimization». Dins *Proceedings of the shared task session of CoNLL 2007*, ps. 933–939. Praga.

HALLIDAY, Michael A.K.; HASAN, Ruqaiya (1976): *Cohesion in English*. Longman, Londres.

HAMMAMI, Souha; BELGUITH, Lamia; HAMADOU, Abdelmajid Ben (2009): «Arabic anaphora resolution: Corpora annotation with coreferential links». *The International Arab Journal of Information Technology*, volum 6(5): ps. 481–489.

HARABAGIU, Sanda; BUNESCU, Razvan; MAIORANO, Steven (2001): «Text and knowledge mining for coreference resolution». Dins *Proceedings of NAACL 2001*, ps. 55–62.

HASLER, Laura; ORASAN, Constantin; NAUMANN, Karin (2006): «NPs for events: Experiments in coreference annotation». Dins *Proceedings of LREC 2006*, ps. 1167–1172. Gènova, Itàlia.

HAYES, Andrew F.; KRIPPENDORFF, Klaus (2007): «Answering the call for a standard reliability measure for coding data». *Communication methods and measures*, volum 1(1): ps. 77–89.

HEIM, Irene (1983): «File change semantics and the familiarity theory of definiteness». Dins *Meaning, Use, and Interpretation of Language* (R. BÄUERLE; C. SCHWARZE; A. VON STECHOW, eds.), ps. 164–189. Mouton de Gruyter, Berlín.

HENDRICKX, Iris; *et al.* (2008): «A coreference corpus and resolution system for Dutch». Dins *Proceedings of LREC 2008*. Marràqueix, Marroc.

HERVÁS, Raquel; FINLAYSON, Mark (2010): «The prevalence of descriptive referring expressions in news and narrative». Dins *Proceedings of ACL 2010 Short Papers*, ps. 49–54. Uppsala, Suècia.

HINRICHS, Erhard; *et al.* (2004): «Recent developments in Linguistic Annotations of the TüBa-D/Z Treebank». Dins *Proceedings of TLT 2004*. Tübingen, Alemanya.

HINRICHS, Erhard W.; FILIPPOVA, Katja; WUNSCH, Holger (2007): «A data-driven approach to pronominal anaphora resolution in German». Dins *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005* (Nicolas NICOLOV; Kalina BONTCHEVA; Galia ANGELOVA; Ruslan MITKOV, eds.), ps. 115–124. John Benjamins, Amsterdam.

HINRICHS, Erhard W.; KÜBLER, Sandra; NAUMANN, Karin (2005): «A unified representation for morphological, syntactic, semantic, and referential annotations». Dins *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, ps. 13–20. Ann Arbor, Michigan.

HIRSCHMAN, Lynette; CHINCHOR, Nancy (1997): «MUC-7 Coreference Task Definition – Version 3.0». Dins *Proceedings of MUC-7*.

HIRST, Graeme J. (1981): *Anaphora in natural language understanding: a survey*. Springer-Verlag, Berlín.

HOBBS, Jerry (1985): «Granularity». Dins *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI 1985)*, ps. 432–435. Los Angeles, Califòrnia.

HOBBS, Jerry R. (1978): «Resolving pronoun references». *Lingua*, volum 44: ps. 311–338.

HOSTE, Véronique; DE PAUW, Guy (2006): «KNACK-2002: A richly annotated corpus of Dutch written text». Dins *Proceedings of LREC 2006*, ps. 1432–1437. Gènova, Itàlia.

HOSTE, Véronique (2005): *Optimization Issues in Machine Learning of Coreference Resolution*. Tesi Doctoral, Universitat d'Anvers, Anvers, Bèlgica.

HOVY, Eduard; *et al.* (2006): «OntoNotes: the 90% solution». Dins *Proceedings of HLT-NAACL 2006*, ps. 57–60. Nova York.

HUBERT, Lawrence; ARABIE, Phipps (1985): «Comparing partitions». *Journal of Classification*, volum 2(1): ps. 193–218.

IDE, Nancy (2000): «Searching annotated language resources in XML: A statement of the problem». Dins *Proceedings of the ACM SIGIR Workshop on XML and Information Retrieval*. Atenes.

IIDA, Ryu; *et al.* (2003): «Incorporating contextual cues in trainable models for coreference resolution». Dins *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, ps. 23–30. Budapest.

— (2007): «Annotating a Japanese text corpus with predicate-argument and coreference relations». Dins *Proceedings of the ACL Workshop on Linguistic Annotation*, ps. 132–139. Praga.

JACKENDOFF, Ray (1983): *Semantics and Cognition*. MIT Press, Cambridge.

— (2002): *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

KABADJOV, Mijail A. (2007): *A comprehensive evaluation of anaphora resolution and discourse-new classification*. Tesi Doctoral, University of Essex, Colchester, Regne Unit.

KAMEYAMA, Megumi (1998): «Intrasentential centering: A case study». Dins *Centering Theory in Discourse* (Marilyn A. WALKER; Aravind K. JOSHI; Ellen F. PRINCE, eds.), ps. 89–112. Oxford University Press, Oxford.

KAMP, Hans (1981): «A theory of truth and semantic representation». Dins *Formal Methods in the Study of Language* (J.A.G. GROENENDIJK; T.M.V. JANSSEN; M.B.J. STOKHOF, eds.), ps. 277–322. Mathematical Centre Tracts 135, Amsterdam.

KARTTUNEN, Lauri (1976): «Discourse referents». Dins *Syntax and Semantics* (J. MCCAWLEY, ed.), volum 7, ps. 363–385. Academic Press, Nova York.

KEHLER, Andrew (1997): «Probabilistic coreference in information extraction». Dins *Proceedings of EMNLP 1997*, ps. 163–173. Providence, Rhode Island.

KEHLER, Andrew; *et al.* (2004): «The (non)utility of predicate-argument frequencies for pronoun interpretation». Dins *Proceedings of NAACL 2004*, ps. 289–296. Boston, Massachusetts.

— (2008): «Coherence and coreference revisited». *Journal of Semantics*, volum 25(1): ps. 1–44.

KENNEDY, Christopher; BOGURAEV, Branimir (1996): «Anaphora for everyone: Pronominal anaphora resolution without a parser». Dins *Proceedings of COLING 1996*, ps. 113–118. Copenhaguen.

KILGARRIFF, Adam (1999): «95% Replicability for manual word sense tagging». Dins *Proceedings of EACL 1999*, ps. 277–278. Bergen, Noruega.

KLENNER, Manfred; AILLOUD, Étienne (2009): «Optimization in coreference resolution is not needed: A nearly-optimal algorithm with intensional constraints». Dins *Proceedings of EACL 2009*, ps. 442–450. Atenes.

KOBDANI, Hamidreza; SCHÜTZE, Hinrich (2010): «SUCRE: A modular system for coreference resolution». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 92–95. Uppsala, Suècia.

KOZAREVA, Zornitsa; HOVY, Eduard (2010): «Learning arguments and supertypes of semantic relations using recursive patterns». Dins *Proceedings of ACL 2010*, ps. 1482–1491. Uppsala, Suècia.

KRAHMER, Emiel (2010): «What computational linguists can learn from psychologists (and vice versa)». *Computational Linguistics*, volum 36(2): ps. 285–294.

KRIPKE, Saul (1977): «Speaker's reference and semantic reference». *Midwest Studies in Philosophy*, volum 2: ps. 255–276.

KRIPPENDORFF, Klaus (2004 [1980]): *Content Analysis: An Introduction to its Methodology*. Sage, Thousand Oaks, Califòrnia, segona edició. Chapter 11.

KUDOH, Taku; MATSUMOTO, Yuji (2000): «Use of support vector learning for chunk identification». Dins *Proceedings of CoNLL 2000 and LLL 2000*, ps. 142–144. Lisboa.

KUČOVÁ, Lucie; HAJIČOVÁ, Eva (2004): «Coreferential relations in the Praga Dependency Treebank». Dins *Proceedings of DAARC 2004*, ps. 97–102. Azores, Portugal.

LAKOFF, George (1987): *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago.

LAPPIN, Shalom; LEASS, Herbert J. (1994): «An algorithm for pronominal anaphora resolution». *Computational Linguistics*, volum 20(4): ps. 535–561.

LASERSOHN, Peter (2000): «*Same*, models and representation». Dins *Proceedings of Semantics and Linguistic Theory 10* (Brendan JACKSON; Tanya MATHEWS, eds.), ps. 83–97. CLC Publications, Cornell, Nova York.

LUO, Xiaoqiang (2005): «On coreference resolution performance metrics». Dins *Proceedings of HLT-EMNLP 2005*, ps. 25–32. Vancouver, Canadà.

— (2007): «Coreference or not: A twin model for coreference resolution». Dins *Proceedings of HLT-NAACL 2007*, ps. 73–80. Rochester, Nova York.

LUO, Xiaoqiang; ZITOUNI, Imed (2005): «Multi-lingual coreference resolution with syntactic features». Dins *Proceedings of HLT-EMNLP 2005*, ps. 660–667. Vancouver, Canadà.

LUO, Xiaoqiang; *et al.* (2004): «A mention-synchronous coreference resolution algorithm based on the Bell tree». Dins *Proceedings of ACL 2004*, ps. 21–26. Barcelona.

MADNANI, Nitin; DORR, Bonnie J. (2010): «Generating phrasal and sentential paraphrases: A survey of data-driven methods». *Computational Linguistics*, volum 36(3): ps. 341–387.

MAGNINI, Bernardo; *et al.* (2006): «I-CAB: the Italian Content Annotation Bank». Dins *Proceedings of LREC 2006*, ps. 963–968. Gènova, Itàlia.

MARKERT, Katja; NISSIM, Malvina (2005): «Comparing knowledge sources for nominal anaphora resolution». *Computational Linguistics*, volum 31(3): ps. 367–401.

MAYOL, Laia; CLARK, Robin (2010): «Pronouns in Catalan: Games of partial information and the use of linguistic resources». *Journal of Pragmatics*, volum 42: ps. 781–799.

MCCALLUM, Andrew; WELLNER, Ben (2005): «Conditional models of identity uncertainty with application to noun coreference». Dins *Advances in Neural Information Processing Systems 17* (Lawrence K. SAUL; Yair WEISS; Léon BOTTOU, eds.), ps. 905–912. MIT Press, Cambridge, Massachusetts.

MCCARTHY, Joseph F.; LEHNERT, Wendy G. (1995): «Using decision trees for coreference resolution». Dins *Proceedings of IJCAI 1995*, ps. 1050–1055. Mont-real, Canadà.

MILIĆEVIĆ, Jasmina (2007): *La paraphrase*. Peter Lang, Berna.

MIRKIN, Shachar; *et al.* (2010): «Recognising entailment within discourse». Dins *Proceedings of COLING 2010*. Pequín.

MITKOV, Ruslan (1998): «Robust pronoun resolution with limited knowledge». Dins *Proceedings of COLING-ACL 1998*, ps. 869–875. Mont-real, Canadà.

— (2002): *Anaphora Resolution*. Longman, Londres.

MITKOV, Ruslan; HALLETT, Catalina (2007): «Comparing pronoun resolution algorithms». *Computational Intelligence*, volum 23(2): ps. 262–97.

MITKOV, Ruslan; *et al.* (2000): «Coreference and anaphora: Developing annotating tools, annotated resources and annotation strategies». Dins *Proceedings of DAARC 2000*, ps. 49–58. Lancaster, Regne Unit.

MORTON, Thomas S. (1999): «Using coreference in question answering». Dins *Proceedings of TREC-8*, ps. 85–89. Gaithersburg, Maryland.

— (2000): «Coreference for NLP applications». Dins *Proceedings of ACL 2000*, ps. 173–180. Hong Kong.

MÜLLER, Christoph (2007): «Resolving *it*, *this* and *that* in unrestricted multi-party dialog». Dins *Proceedings of ACL 2007*, ps. 816–823. Praga.

MÜLLER, Christoph; RAPP, Stefan; STRUBE, Michael (2002): «Applying co-training to reference resolution». Dins *Proceedings of ACL 2002*, ps. 352–359. Filadèlfia, Pennsilvània.

MÜLLER, Christoph; STRUBE, Michael (2006): «Multi-level annotation of linguistic data with MMAX2». Dins *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods* (Sabine BRAUN; Kurt KOHN; Joybrato MUKHERJEE, eds.), ps. 197–214. Peter Lang, Frankfurt.

NAVARRETTA, Costanza (2004): «Resolving individual and abstract anaphora in texts and dialogues». Dins *Proceedings of COLING 2004*, ps. 233–239. Ginebra, Suïssa.

— (2007): «A contrastive analysis of abstract anaphora in Danish, English and Italian». Dins *Proceedings of DAARC 2007*, ps. 103–109. Lagos, Portugal.

— (2009*a*): «Automatic recognition of the function of singular neuter pronouns in texts and spoken data». Dins *Anaphora Processing and Applications (DAARC 2009)* (S. Lalitha DEVI; A. BRANCO; R. MITKOV, eds.), *LNAI*, volum 5847, ps. 15–28. Springer-Verlag, Berlín / Heidelberg.

— (2009*b*): «Co-referential chains and discourse topic shifts in parallel and comparable corpora». *Procesamiento del Lenguaje Natural*, volum 42: ps. 105–112.

NAVARRO, Borja (2007): *Metodología, construcción y explotación de corpus anotados semántica y anafóricamente*. Tesi Doctoral, Universitat d'Alacant, Alacant, Espanya.

NG, Vincent (2003): «Machine learning for coreference resolution: Recent successes and future challenges». Technical report CUL.CIS/TR2003-1918, Cornell University, Nova York.

— (2004): «Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization». Dins *Proceedings of ACL 2004*, ps. 151–158. Barcelona.

— (2005): «Machine learning for coreference resolution: from local classification to global ranking». Dins *Proceedings of ACL 2005*, ps. 157–164. Ann Arbor, Michigan.

— (2007): «Shallow semantics for coreference resolution». Dins *Proceedings of IJCAI 2007*, ps. 1689–1694. Hyderabad, Índia.

— (2008): «Unsupervised models for coreference resolution». Dins *Proceedings of EMNLP 2008*, ps. 640–649. Honolulu, Hawaii.

— (2009): «Graph-cut-based anaphoricity determination for coreference resolution». Dins *Proceedings of NAACL-HLT 2009*, ps. 575–583. Boulder, Colorado.

— (2010): «Supervised noun phrase coreference research: The first fifteen years». Dins *Proceedings of ACL 2010*, ps. 1396–1411. Uppsala, Suècia.

NG, Vincent; CARDIE, Claire (2002*a*): «Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution». Dins *Proceedings of COLING 2002*, ps. 1–7. Taipei.

— (2002*b*): «Improving machine learning approaches to coreference resolution». Dins *Proceedings of ACL 2002*, ps. 104–111. Filadèlfia, Pennsilvània.

NICOLAE, Cristina; NICOLAE, Gabriel (2006): «BestCut: a graph algorithm for coreference resolution». Dins *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, ps. 275–283. Sydney, Austràlia.

NICOLAE, Cristina; NICOLAE, Gabriel; ROBERTS, Kirk (2010): «C-3: Coherence and coreference corpus». Dins *Proceedings of LREC 2010*, ps. 136–143. Valletta, Malta.

NICOLOV, Nicolas; SALVETTI, Franco; IVANOVA, Steliana (2008): «Sentiment analysis: Does coreference matter?» Dins *Proceedings of the Symposium on Affective Language in Human and Machine*. Aberdeen, Regne Unit.

NILSSON, Kristina (2010): *Hybrid Methods for Coreference Resolution in Swedish*. Tesi Doctoral, Universitat d'Estocolm, Estocolm.

VAN NOORD, Gertjan; SCHUURMAN, Ineke; VANDEGHINSTE, Vincent (2006): «Syntactic annotation of large corpora in STEVIN». Dins *Proceedings of LREC 2006*, ps. 1811–1814. Gènova, Itàlia.

NUNBERG, Geoffrey (1984): «Individuation in context». Dins *Proceedings of the 2nd West Coast Conference on Formal Linguistics (WCCFL 2)*, ps. 203–217. Stanford, Califòrnia.

ORASAN, Constantin (2003): «PALinkA: A highly customisable tool for discourse annotation». Dins *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, ps. 39–43. Sapporo, Japó.

ORASAN, Constantin; *et al.* (2008): «Anaphora Resolution Exercise: An overview». Dins *Proceedings of LREC 2008*, ps. 2801–2805. Marràqueix, Marroc.

PALOMAR, Manuel; *et al.* (2001): «An algorithm for anaphora resolution in Spanish texts». *Computational Linguistics*, volum 27(4): ps. 545–567.

PASSONNEAU, Rebecca (2004): «Computing reliability for coreference annotation». Dins *Proceedings of LREC 2004*, ps. 1503–1506. Lisboa.

— (2006): «Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation». Dins *Proceedings of LREC 2006*, ps. 831–836. Gènova, Itàlia.

PEÑAS, Anselmo; HOVY, Eduard (2010): «Semantic enrichment of text with background knowledge». Dins *Proceedings of the NAACL First International Workshop on Formalisms and Methodology for Learning by Reading*, ps. 15–23. Los Angeles, Califòrnia.

POESIO, Massimo (2004*a*): «Discourse annotation and semantic annotation in the GNOME corpus». Dins *Proceedings of the ACL Workshop on Discourse Annotation*, ps. 72–79. Barcelona.

— (2004*b*): «The MATE/GNOME proposals for anaphoric annotation, revisited». Dins *Proceedings of the 5th SIGdial Workshop at HLT-NAACL 2004*, ps. 154–162. Boston, Massachusetts.

POESIO, Massimo; ARTSTEIN, Ron (2005): «The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account». Dins *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, ps. 76–83. Ann Arbor, Michigan.

— (2008): «Anaphoric annotation in the ARRAU corpus». Dins *Proceedings of LREC 2008*. Marràqueix, Marroc.

POESIO, Massimo; KRUSCHWITZ, Udo; CHAMBERLAIN, Jon (2008): «ANAWIKI: Creating anaphorically annotated resources through Web cooperation». Dins *Proceedings of LREC 2008*, ps. 2352–2355. Marràqueix, Marroc.

POESIO, Massimo; PONZETTO, Simone Paolo; VERSLEY, Yannick (en preparació): «Computational models of anaphora resolution: A survey». *Linguistic Issues in Language Technology*.

POESIO, Massimo; URYUPINA, Olga; VERSLEY, Yannick (2010): «Creating a coreference resolution system for Italian». Dins *Proceedings of LREC 2010*, ps. 713–716. Valletta, Malta.

POESIO, Massimo; VIEIRA, Renata (1998): «A corpus-based investigation of definite description use». *Computational Linguistics*, volum 24(2): ps. 183–216.

POESIO, Massimo; *et al.* (2004*a*): «Centering: a parametric theory and its instantiations». *Computational Linguistics*, volum 30(3): ps. 309–363.

— (2004*b*): «The VENEX corpus of anaphora and deixis in spoken and written Italian». Manuscript. Available online at `http://cswww.essex.ac.RegneUnit/staff/poesio/publications/VENEX04.pdf`.

PONZETTO, Simone Paolo; STRUBE, Michael (2006): «Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution». Dins *Proceedings of HLT-NAACL 2006*, ps. 192–199. Nova York.

POON, Hoifung; DOMINGOS, Pedro (2008): «Joint unsupervised coreference resolution with Markov logic». Dins *Proceedings of EMNLP 2008*, ps. 650–659. Honolulu, Hawaii.

POON, Hoifung; *et al.* (2010): «Machine Reading at the University of Washington». Dins *Proceedings of the NAACL-HLT First International Workshop on Formalisms and Methodology for Learning by Reading*, ps. 87–95. Los Angeles, Califòrnia.

POPESCU-BELIS, Andrei (2000): «Évaluation numérique de la résolution de la référence: critiques et propositions». *T.A.L.: Traitement automatique de la langue*, volum 40(2): ps. 117–146.

POPESCU-BELIS, Andrei; ROBBA, Isabelle; SABAH, Gérard (1998): «Reference resolution beyond coreference: a conceptual frame and its application». Dins *Proceedings of COLING-ACL 1998*, ps. 1046–1052. Mont-real, Canadà.

POPESCU-BELIS, Andrei; *et al.* (2004): «Online evaluation of coreference resolution». Dins *Proceedings of LREC 2004*, ps. 1507–1510. Lisboa.

PRADHAN, Sameer S.; *et al.* (2007*a*): «OntoNotes: A unified relational semantic representation». Dins *Proceedings of ICSC 2007*, ps. 517–526. Irvine, Califòrnia.

— (2007*b*): «Unrestricted coreference: Identifying entities and events in OntoNotes». Dins *Proceedings of ICSC 2007*, ps. 446–453. Irvine, Califòrnia.

PRINCE, Ellen F. (1981): «Toward a taxonomy of given-new information». Dins *Radical Pragmatics* (Peter COLE, ed.), ps. 223–256. Academic Press, Nova York.

QUINLAN, Ross (1993): *C4.5: Program for Machine Learning*. Morgan Kaufmann, San Francisco, Califòrnia.

RAHMAN, Altaf; NG, Vincent (2009): «Supervised models for coreference resolution». Dins *Proceedings of EMNLP 2009*, ps. 968–977. Suntec, Singapur.

RAND, William M. (1971): «Objective criteria for the evaluation of clustering methods». *Journal of the American Statistical Association*, volum 66(336): ps. 846–850.

RECASENS, Marta (2008): «Discourse deixis and coreference: Evidence from AnCora». Dins *Proceedings of the Second Workshop on Anaphora Resolution (WAR II)*, *NEALT Proceedings Series*, volum 2, ps. 73–82. Bergen, Noruega.

— (2009): «A chain-starting classifier of definite NPs in Spanish». Dins *Proceedings of the EACL Student Research Workshop (EACL 2009)*, ps. 46–53. Atenes.

RECASENS, Marta; HOVY, Eduard (2009): «A deeper look into features for coreference resolution». Dins *Anaphora Processing and Applications (DAARC 2009)* (S. Lalitha DEVI; A. BRANCO; R. MITKOV, eds.), *LNAI*, volum 5847, ps. 29–42. Springer-Verlag, Berlín.

— (2010): «Coreference resolution across corpora: Languages, coding schemes, and preprocessing information». Dins *Proceedings of ACL 2010*, ps. 1423–1432. Uppsala, Suècia.

— (en premsa): «BLANC: Implementing the Rand index for coreference evaluation». *Natural Language Engineering*.

RECASENS, Marta; HOVY, Eduard; MARTÍ, M. Antònia (2010*a*): «A typology of near-identity relations for coreference (NIDENT)». Dins *Proceedings of LREC 2010*, ps. 149–156. Valletta, Malta.

RECASENS, Marta; MARTÍ, M. Antònia (2010): «AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan». *Language Resources and Evaluation*, volum 44(4): ps. 315–345.

RECASENS, Marta; MARTÍ, M. Antònia; TAULÉ, Mariona (2009*a*): «First-mention definites: More than exceptional cases». Dins *The Fruits of Empirical Linguistics* (S. FEATHERSTON; S. WINKLER, eds.), volum 2, ps. 217–237. Mouton de Gruyter, Berlín.

RECASENS, Marta; *et al.* (2009*b*): «SemEval-2010 Task 1: Coreference resolution in multiple languages». Dins *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, ps. 70–75. Boulder, Colorado.

— (2010*b*): «SemEval-2010 Task 1: Coreference resolution in multiple languages». Dins *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, ps. 1–8. Uppsala, Suècia.

RICH, Elaine; LUPERFOY, Susann (1988): «An architecture for anaphora resolution». Dins *Proceedings of ANLP 1988*, ps. 18–24. Austin, Texas.

RODRÍGUEZ, Kepa Joseba; *et al.* (2010): «Anaphoric annotation of Wikipedia and blogs in the Live Memories Corpus». Dins *Proceedings of LREC 2010*, ps. 157–163. Valletta, Malta.

RUPPENHOFER, Josef; SPORLEDER, Caroline; MORANTE, Roser (2010): «SemEval-2010 Task 10: Linking events and their participants in discourse». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 45–50. Uppsala, Suècia.

RUSSELL, Bertrand (1905): «On denoting». *Mind*, volum 15: ps. 479–493.

SAPENA, Emili; PADRÓ, Lluís; TURMO, Jordi (2010): «RelaxCor: A global relaxation labeling approach to coreference resolution for the Semeval-2010 Coreference Task». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 88–91. Uppsala, Suècia.

SCHMID, Helmut (1995): «Improvements in part-of-speech tagging with an application to German». Dins *Proceedings of the EACL SIGDAT Workshop*, ps. 47–50. Dublín.

SCHMID, Helmut; LAWS, Florian (2008): «Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging». Dins *Proceedings of COLING 2008*, ps. 777–784. Manchester, Regne Unit.

SHINYAMA, YusRegne Unite; SEKINE, Satoshi (2003): «Paraphrase acquisition for information extraction». Dins *Proceedings of the ACL 2nd International Workshop on Paraphrasing (IWP 2003)*, ps. 65–71. Sapporo, Japó.

SIEGEL, Sidney; CASTELLAN, N. John (1988): *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill, Nova York, segona edició. Chapter 9.8.

SOLÀ, Joan (ed.) (2002): *Gramàtica del català contemporani*. Empúries, Barcelona.

SOON, Wee M.; NG, Hwee T.; LIM, Daniel C. Y. (2001): «A machine learning approach to coreference resolution of noun phrases». *Computational Linguistics*, volum 27(4): ps. 521–544.

STEDE, Manfred (2004): «The Potsdam Commentary Corpus». Dins *Proceedings of the ACL Workshop on Discourse Annotation*, ps. 96–102. Barcelona.

STEINBERGER, Josef; *et al.* (2007): «Two uses of anaphora resolution in summarization». *Information Processing and Management: an International Journal*, volum 43(6): ps. 1663–1680.

STEVENSON, Rosemary; CRAWLEY, Rosalind; KLEINMAN, David (1994): «Thematic roles, focus and the representation of events». *Language and Cognitive Processes*, volum 9: ps. 519–548.

STOYANOV, Veselin; *et al.* (2009): «Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art». Dins *Proceedings of ACL-IJCNLP 2009*, ps. 656–664. Suntec, Singapur.

— (2010): «Coreference resolution with Reconcile». Dins *Proceedings of ACL 2010 Short Papers*, ps. 156–161. Uppsala, Suècia.

STRUBE, Michael; MÜLLER, Christoph (2003): «A machine learning approach to pronoun resolution in spoken dialogue». Dins *Proceedings of ACL 2003*, ps. 168–175. Sapporo, Japó.

STRUBE, Michael; RAPP, Stefan; MÜLLER, Christoph (2002): «The influence of minimum edit distance on reference resolution». Dins *Proceedings of ACL-EMNLP 2002*, ps. 312–319.

TABOADA, Maite (2008): «Reference, centers and transitions in spoken Spanish». Dins *Reference and Reference Processing* (J. GUNDEL; N. HEDBERG, eds.), ps. 176–215. Oxford University Press, Oxford.

TAULÉ, Mariona; MARTÍ, M. Antònia; RECASENS, Marta (2008): «AnCora: Multilevel annotated corpora for Catalan and Spanish». Dins *Proceedings of LREC 2008*, ps. 96–101. Marràqueix, Marroc.

TETREAULT, Joel (1999): «Analysis of syntax-based pronoun resolution methods». Dins *Proceedings of ACL 1999*, ps. 602–605. College Park, Maryland.

— (2001): «A corpus-based evaluation of centering and pronoun resolution». *Computational Linguistics*, volum 27(4): ps. 507–520.

TJONG KIM SANG, Erik F.; DE MEULDER, Fien (2003): «Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition». Dins *Proceedings of CoNLL 2003* (Walter DAELEMANS; Miles OSBORNE, eds.), ps. 142–147. Edmonton, Canada.

URYUPINA, Olga (2003): «High-precision identification of discourse-new and unique noun phrases». Dins *Proceedings of the ACL 2003 Student Workshop*, ps. 80–86. Sapporo, Japó.

— (2004): «Linguistically motivated sample selection for coreference resolution». Dins *Proceedings of DAARC 2004*. Azores, Portugal.

— (2006): «Coreference resolution with and without linguistic knowledge». Dins *Proceedings of LREC 2006*, ps. 893–898. Gènova, Itàlia.

— (2007): *Knowledge Acquisition for Coreference Resolution*. Tesi Doctoral, Universitat de Saarland, Saarbrücken, Alemanya.

— (2008): «Error analysis for learning-based coreference resolution». Dins *Proceedings of LREC 2008*, ps. 1914–1919. Marràqueix, Marroc.

— (2010): «Corry: A system for coreference resolution». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 100–103. Uppsala, Suècia.

VAN DEEMTER, Kees; KIBBLE, Rodger (2000): «On coreferring: Coreference in MUC and related annotation schemes». *Computational Linguistics*, volum 26(4): ps. 629–637.

VERSLEY, Yannick (2007): «Antecedent selection techniques for high-recall coreference resolution». Dins *Proceedings of EMNLP-CoNLL 2007*, ps. 496–505. Praga.

— (2008): «Vagueness and referential ambiguity in a large-scale annotated corpus». *Research on Language and Computation*, volum 6: ps. 333–353.

VERSLEY, Yannick; *et al.* (2008): «BART: A modular toolkit for coreference resolution». Dins *Proceedings of LREC 2008*, ps. 962–965. Marràqueix, Marroc.

VICEDO, Jose L.; FERRÁNDEZ, Antonio (2006): «Coreference in Q&A». Dins *Advances in Open Domain Question Answering* (Tomek STRZALKOWSKI; Sanda HARABAGIU, eds.), *Text, Speech and Language Technology*, volum 32, ps. 71–96. Springer-Verlag, Berlín.

VIEIRA, Renata; POESIO, Massimo (2000): «An empirically-based system for processing definite descriptions». *Computational Linguistics*, volum 26(4): ps. 539–593.

VILA, Marta; *et al.* (2010): «ClInt: a bilingual Spanish-Catalan spoken corpus of clinical interviews». *Procesamiento del Lenguaje Natural*, volum 45: ps. 105–111.

VILAIN, Marc; *et al.* (1995): «A model-theoretic coreference scoring scheme». Dins *Proceedings of MUC-6*, ps. 45–52.

VINH, Nguyen Xuan; EPPS, Julien; BAILEY, James (2009): «Information theoretic measures for clusterings comparison: Is a correction for chance necessary?» Dins *Proceedings of ICML 2009*, ps. 577–584. Mont-real, Canadà.

WEBBER, Bonnie Lynn (1979): *A Formal Approach to Discourse Anaphora*. Garland Press, Nova York.

— (1988): «Discourse deixis: reference to discourse segments». Dins *Proceedings of ACL 1988*, ps. 113–122. Buffalo, Nova York.

WICK, Michael; MCCALLUM, Andrew (2009): «Advances in learning and inference for partition-wise models of coreference resolution». Report tècnic UM-CS-2009-028, University of Massachusetts, Amherst, Massachusetts.

WICK, Michael; *et al.* (2009): «An entity based model for coreference resolution». Dins *Proceedings of SDM 2009*, ps. 365–376. Reno, Nevada.

WINTNER, Shuly (2009): «What science underlies Natural Language Engineering?» *Computational Linguistics*, volum 35(4): ps. 641–644.

WITTGENSTEIN, Ludwig (1953): *Philosophical Investigations*. Blackwell, Oxford.

YANG, Xiaofeng; SU, Jian (2007): «Coreference resolution using semantic relatedness information from automatically discovered patterns». Dins *Proceedings of ACL 2007*, ps. 525–535. Praga.

YANG, Xiaofeng; *et al.* (2003): «Coreference resolution using competition learning approach». Dins *Proceedings of ACL 2003*, ps. 176–183. Sapporo, Japó.

— (2004): «Improving pronoun resolution by incorporating coreferential information of candidates». Dins *Proceedings of ACL 2004*, ps. 127–134. Barcelona.

— (2008): «An entity-mention model for coreference resolution with inductive logic programming». Dins *Proceedings of ACL-HLT 2008*, ps. 843–851. Columbus, Ohio.

ZAENEN, Annie (2006): «Mark-up barking up the wrong tree». *Computational Linguistics*, volum 32(4): ps. 577–580.

ZHEKOVA, Desislava; KÜBLER, Sandra (2010): «UBIU: A language-independent system for coreference resolution». Dins *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, ps. 96–99. Uppsala, Suècia.

APÈNDIXS

APÈNDIX A

Sortides de sistemes

Aquest apèndix recull, per a dos documents d'OntoNotes, les sortides de coreferència de diferents versions de CISTELL (ENCAIX FORT, ENCAIX SUPERFORT, ENCAIX ÒPTIM, ENCAIX FEBLE; tal com s'explica a l'apartat 4.3.2) juntament amb els *baselines* de TOTES UNÀRIES i MATEIX NUCLI. També s'inclouen per al primer document les sortides dels sis sistemes que van participar al SemEval (apartat 6.4). Les mencions que corefereixen comparteixen el mateix subíndex.

A.1 Fitxer d'OntoNotes nbc_0030

The nation's highest court will take up the case next week. That development may not be as significant as it seems. Joining me now is law professor Rick Pildes, a consultant to NBC News. Could a decision from the U.S. Supreme Court settle this case once and for all? <TURN> At this stage, any decision from the U.S. Supreme Court is almost certainly not going to provide a final resolution of this election dispute. Indeed, the issue is so narrow now before the Supreme Court that whichever way the court rules, it will likely have only the most marginal impact on what's going on in Florida. Even if the Bush campaign prevails before the Supreme Court, it simply means we will move more quickly into the contest phase of the litigation or the next stage of the litigation. <TURN> But you believe the fact that the U.S. Supreme Court just decided to hear this case is a partial victory for both Bush and Gore. <TURN> It is a partial victory for both sides. For the last two

weeks, the central constitutional argument the Bush campaign has been making to the federal courts is, stop these manual recounts now, they violate the Constitution. The U.S. Supreme Court refused to hear that part of the case, agreeing with all the other federal judges who have unanimously held that this is not the proper time for federal court intervention. So in that sense, a victory for the Gore campaign. For the Bush campaign, a victory in the willingness of the Supreme Court to play some role in overseeing the Florida system and the Florida judicial decision making process. <TURN> Whatever the Supreme Court decides this time, you say this case could come back before the U.S. Supreme Court again? <TURN> John, if the Supreme Court of the United States is to play a final and decisive role in this dispute, that role is going to come at the end of the Florida judicial process, not at this stage. <TURN> Law professor Rick Pildes, thank you.

1. **ESTÀNDARD DE REFERÈNCIA**
   [[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_4$ seems. Joining [me]$_5$ now is [law professor Rick Pildes, a consultant to [NBC News]$_7$]$_6$. Could [a decision from [the U.S. Supreme Court]$_0$]$_8$ settle [this case]$_2$ once and for [all]$_9$? <TURN> At [this stage]$_{10}$, [any decision from [the U.S. Supreme Court]$_0$]$_{11}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{13}$]$_{12}$. Indeed, [the issue]$_{14}$ is so narrow now before [the Supreme Court]$_0$ that whichever way [the court]$_0$ rules, [it]$_{15}$ will likely have [only the most marginal impact on what's going on in [Florida]$_{17}$]$_{16}$. Even if [the [Bush]$_{19}$ campaign]$_{18}$ prevails before [the Supreme Court]$_0$, [it]$_{20}$ simply means [we]$_{21}$ will move more quickly into [the contest phase of [the litigation]$_{23}$ or [the next stage of [the litigation]$_{23}$]$_{24}$]$_{22}$. <TURN> But [you]$_6$ believe [the fact that [the U.S. Supreme Court]$_0$ just decided to hear [this case]$_2$]$_{25}$ is [a partial victory for [both [Bush]$_{19}$ and [Gore]$_{28}$]$_{27}$]$_{26}$. <TURN> [It]$_{25}$ is [a partial victory for [both sides]$_{30}$]$_{29}$. For [the last two weeks]$_{31}$, [the central constitutional argument [the [Bush]$_{19}$ campaign]$_{18}$ has been making to [the federal courts]$_{33}$]$_{32}$ is, stop [these manual recounts]$_{34}$ now, [they]$_{34}$ violate [the Constitution]$_{35}$. [The U.S. Supreme Court]$_0$ refused to hear [that part of [the case]$_2$]$_{32}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{37}$ is not [the proper time for [federal court intervention]$_{39}$]$_{38}$]$_{36}$. So in [that sense]$_{40}$, [a victory for [the [Gore]$_{28}$ campaign]$_{42}$]$_{41}$. For [the [Bush]$_{19}$ campaign]$_{18}$, [a victory in [the willingness of [the Supreme Court]$_0$ to play [some role in overseeing [the [Florida]$_{17}$ system and [the [Florida]$_{17}$ judicial decision making process]$_{47}$]$_{46}$]$_{45}$]$_{44}$]$_{43}$. <TURN> Whatever [the Supreme Court]$_0$ decides [this time]$_{48}$, [you]$_6$ say [this case]$_2$ could come back before [the U.S. Supreme Court]$_0$ again? <TURN> [John]$_5$, if [the Supreme Court of [the United States]$_1$]$_0$ is to play [a final and decisive role in [this dispute]$_{13}$]$_{49}$, [that role]$_{49}$ is going to come at [the end of [the [Florida]$_{17}$ judicial

process]$_{47}$]$_{50}$, not at [this stage]$_{10}$. <TURN> [Law professor Rick Pildes]$_{6}$, thank [you]$_{6}$.

2. **BASELINE: TOTES UNÀRIES**
[[The nation's]$_{1}$ highest court]$_{0}$ will take up [the case]$_{2}$ [next week]$_{3}$. [That development]$_{4}$ may not be as significant as [it]$_{5}$ seems. Joining [me]$_{6}$ now is [law professor Rick Pildes, a consultant to [NBC News]$_{8}$]$_{7}$. Could [a decision from [the U.S. Supreme Court]$_{10}$]$_{9}$ settle [this case]$_{11}$ once and for [all]$_{12}$? <TURN> At [this stage]$_{13}$, [any decision from [the U.S. Supreme Court]$_{15}$]$_{14}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{17}$]$_{16}$. Indeed, [the issue]$_{18}$ is so narrow now before [the Supreme Court]$_{19}$ that whichever way [the court]$_{20}$ rules, [it]$_{21}$ will likely have [only the most marginal impact on what's going on in [Florida]$_{23}$]$_{22}$. Even if [the [Bush]$_{25}$ campaign]$_{24}$ prevails before [the Supreme Court]$_{26}$, [it]$_{27}$ simply means [we]$_{28}$ will move more quickly into [the contest phase of [the litigation]$_{30}$ or [the next stage of [the litigation]$_{32}$]$_{31}$]$_{29}$. <TURN> But [you]$_{33}$ believe [the fact that [the U.S. Supreme Court]$_{35}$ just decided to hear [this case]$_{36}$]$_{34}$ is [a partial victory for [both [Bush]$_{39}$ and [Gore]$_{40}$]$_{38}$]$_{37}$. <TURN> [It]$_{41}$ is [a partial victory for [both sides]$_{43}$]$_{42}$. For [the last two weeks]$_{44}$, [the central constitutional argument [the [Bush]$_{47}$ campaign]$_{46}$ has been making to [the federal courts]$_{48}$]$_{45}$ is, stop [these manual recounts]$_{49}$ now, [they]$_{50}$ violate [the Constitution]$_{51}$. [The U.S. Supreme Court]$_{52}$ refused to hear [that part of [the case]$_{54}$]$_{53}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{56}$ is not [the proper time for [federal court intervention]$_{58}$]$_{57}$]$_{55}$. So in [that sense]$_{59}$, [a victory for [the [Gore]$_{62}$ campaign]$_{61}$]$_{60}$. For [the [Bush]$_{64}$ campaign]$_{63}$, [a victory in [the willingness of [the Supreme Court]$_{67}$ to play [some role in overseeing [the [Florida]$_{70}$ system and [the [Florida]$_{72}$ judicial decision making process]$_{71}$]$_{69}$]$_{68}$]$_{66}$]$_{65}$. <TURN> Whatever [the Supreme Court]$_{73}$ decides [this time]$_{74}$, [you]$_{75}$ say [this case]$_{76}$ could come back before [the U.S. Supreme Court]$_{77}$ again? <TURN> [John]$_{78}$, if [the Supreme Court of [the United States]$_{80}$]$_{79}$ is to play [a final and decisive role in [this dispute]$_{82}$]$_{81}$, [that role]$_{83}$ is going to come at [the end of [the [Florida]$_{86}$ judicial process]$_{85}$]$_{84}$, not at [this stage]$_{87}$. <TURN> [Law professor Rick Pildes]$_{88}$, thank [you]$_{89}$.

3. **BASELINE: MATEIX NUCLI**
[[The nation's]$_{1}$ highest court]$_{0}$ will take up [the case]$_{2}$ [next week]$_{3}$. [That development]$_{4}$ may not be as significant as [it]$_{5}$ seems. Joining [me]$_{6}$ now is [law professor Rick Pildes, a consultant to [NBC News]$_{8}$]$_{7}$. Could [a decision from [the U.S. Supreme Court]$_{10}$]$_{9}$ settle [this case]$_{2}$ once and for [all]$_{11}$? <TURN> At [this stage]$_{12}$, [any decision from [the U.S. Supreme Court]$_{10}$]$_{9}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{14}$]$_{13}$. Indeed, [the issue]$_{15}$ is so narrow now before [the

Supreme Court]$_{10}$ that whichever way [the court]$_0$ rules, [it]$_{16}$ will likely have [only the most marginal impact on what's going on in [Florida]$_{51}$]$_{17}$. Even if [the Bush campaign]$_{18}$ prevails before [the Supreme Court]$_{10}$, [it]$_{19}$ simply means [we]$_{20}$ will move more quickly into [the contest phase of [the litigation]$_{22}$ or [the next stage of [the litigation]$_{22}$]$_{23}$]$_{21}$. <TURN> But [you]$_{24}$ believe [the fact that [the U.S. Supreme Court]$_{10}$ just decided to hear [this case]$_2$]$_{25}$ is [a partial victory for [both Bush and Gore]$_{27}$]$_{26}$. <TURN> [It]$_{28}$ is [a partial victory for [both sides]$_{29}$]$_{26}$. For [the last two weeks]$_{30}$, [the central constitutional argument [the Bush campaign]$_{18}$ has been making to [the federal courts]$_{32}$]$_{31}$ is, stop [these manual recounts]$_{33}$ now, [they]$_{34}$ violate [the Constitution]$_{35}$. [The U.S. Supreme Court]$_{10}$ refused to hear [that part of [the case]$_2$]$_{36}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{38}$ is not [the proper time for [federal court intervention]$_{40}$]$_{39}$]$_{37}$. So in [that sense]$_{41}$, [a victory for [the Gore campaign]$_{18}$]$_{26}$. For [the Bush campaign]$_{18}$, [a victory in [the willingness of [the Supreme Court]$_{10}$ to play [some role in overseeing [the Florida system and [the Florida judicial decision making process]$_{45}$]$_{44}$]$_{43}$]$_{42}$]$_{26}$. <TURN> Whatever [the Supreme Court]$_{10}$ decides [this time]$_{30}$, [you]$_{46}$ say [this case]$_2$ could come back before [the U.S. Supreme Court]$_{10}$ again? <TURN> [John]$_{47}$, if [the Supreme Court of [the United States]$_{48}$]$_{10}$ is to play [a final and decisive role in [this dispute]$_{14}$]$_{43}$, [that role]$_{43}$ is going to come at [the end of [the Florida judicial process]$_{45}$]$_{49}$, not at [this stage]$_{13}$. <TURN> [Law professor Rick Pildes]$_7$, thank [you]$_{50}$.

4. **ENCAIX FORT**

[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_2$ seems. Joining [me]$_5$ now is [law professor Rick Pildes, a consultant to [NBC News]$_7$]$_6$. Could [a decision from [the U.S. Supreme Court]$_0$]$_5$ settle [this case]$_2$ once and for [all]$_8$? <TURN> At [this stage]$_9$, [any decision from [the U.S. Supreme Court]$_0$]$_5$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{11}$]$_{10}$. Indeed, [the issue]$_{12}$ is so narrow now before [the Supreme Court]$_0$ that whichever way [the court]$_0$ rules, [it]$_2$ will likely have [only the most marginal impact on what's going on in [Florida]$_{14}$]$_{13}$. Even if [the Bush campaign]$_{15}$ prevails before [the Supreme Court]$_0$, [it]$_2$ simply means [we]$_{16}$ will move more quickly into [the contest phase of [the litigation]$_{18}$ or [the next stage of [the litigation]$_{18}$]$_9$]$_{17}$. <TURN> But [you]$_{14}$ believe [the fact that [the U.S. Supreme Court]$_0$ just decided to hear [this case]$_{19}$]$_7$ is [a partial victory for [both Bush and Gore]$_{21}$]$_{20}$. <TURN> [It]$_{22}$ is [a partial victory for [both sides]$_{23}$]$_{20}$. For [the last two weeks]$_{24}$, [the central constitutional argument [the Bush campaign]$_{15}$ has been making to [the federal courts]$_{26}$]$_{25}$ is, stop [these manual recounts]$_{27}$ now, [they]$_{27}$ violate [the Constitution]$_{28}$. [The U.S. Supreme Court]$_0$ refused to hear [that part of [the case]$_2$]$_{29}$, agreeing with [all the other federal judges who ha-

ve unanimously held that [this]$_{19}$ is not [the proper time for [federal court intervention]$_{32}$]$_{31}$]$_{30}$. So in [that sense]$_{33}$, [a victory for [the Gore campaign]$_{15}$]$_{34}$. For [the Bush campaign]$_{15}$, [a victory in [the willingness of [the Supreme Court]$_0$ to play [some role in overseeing [the Florida system and [the Florida judicial decision making process]$_{38}$]$_{37}$]$_{36}$]$_{35}$]$_{34}$. <TURN> Whatever [the Supreme Court]$_0$ decides [this time]$_{31}$, [you]$_{14}$ say [this case]$_2$ could come back before [the U.S. Supreme Court]$_{22}$ again? <TURN> [John]$_{39}$, if [the Supreme Court of [the United States]$_1$]$_0$ is to play [a final and decisive role in [this dispute]$_{11}$]$_{40}$, [that role]$_{36}$ is going to come at [the end of [the Florida judicial process]$_{38}$]$_{41}$, not at [this stage]$_9$. <TURN> [Law professor Rick Pildes]$_6$, thank [you]$_6$.

5. **ENCAIX SUPERFORT**

[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_2$ seems. Joining [me]$_5$ now is [law professor Rick Pildes, a consultant to [NBC News]$_7$]$_6$. Could [a decision from [the U.S. Supreme Court]$_8$]$_5$ settle [this case]$_9$ once and for [all]$_{10}$? <TURN> At [this stage]$_{11}$, [any decision from [the U.S. Supreme Court]$_{13}$]$_{12}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{15}$]$_{14}$. Indeed, [the issue]$_{16}$ is so narrow now before [the Supreme Court]$_8$ that whichever way [the court]$_0$ rules, [it]$_2$ will likely have [only the most marginal impact on what's going on in [Florida]$_{18}$]$_{17}$. Even if [the Bush campaign]$_{19}$ prevails before [the Supreme Court]$_{20}$, [it]$_2$ simply means [we]$_{21}$ will move more quickly into [the contest phase of [the litigation]$_{23}$ or [the next stage of [the litigation]$_{23}$]$_{11}$]$_{22}$. <TURN> But [you]$_{18}$ believe [the fact that [the U.S. Supreme Court]$_{24}$ just decided to hear [this case]$_{25}$]$_7$ is [a partial victory for [both Bush and Gore]$_{27}$]$_{26}$. <TURN> [It]$_2$ is [a partial victory for [both sides]$_{29}$]$_{28}$. For [the last two weeks]$_{30}$, [the central constitutional argument [the Bush campaign]$_{32}$ has been making to [the federal courts]$_{33}$]$_{31}$ is, stop [these manual recounts]$_{34}$ now, [they]$_{34}$ violate [the Constitution]$_{35}$. [The U.S. Supreme Court]$_{36}$ refused to hear [that part of [the case]$_2$]$_{37}$, agreeing with [all the other federal judges who have unanimously held that [this]$_9$ is not [the proper time for [federal court intervention]$_{40}$]$_{39}$]$_{38}$. So in [that sense]$_{41}$, [a victory for [the Gore campaign]$_{19}$]$_{42}$. For [the Bush campaign]$_{43}$, [a victory in [the willingness of [the Supreme Court]$_{46}$ to play [some role in overseeing [the Florida system and [the Florida judicial decision making process]$_{49}$]$_{48}$]$_{47}$]$_{45}$]$_{44}$. <TURN> Whatever [the Supreme Court]$_{50}$ decides [this time]$_{51}$, [you]$_{18}$ say [this case]$_{52}$ could come back before [the U.S. Supreme Court]$_{20}$ again? <TURN> [John]$_{53}$, if [the Supreme Court of [the United States]$_1$]$_{54}$ is to play [a final and decisive role in [this dispute]$_{15}$]$_{55}$, [that role]$_{56}$ is going to come at [the end of [the Florida judicial process]$_{58}$]$_{57}$, not at [this stage]$_{59}$. <TURN> [Law professor Rick Pildes]$_6$, thank [you]$_6$.

6. **ENCAIX ÒPTIM**
[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_2$ seems. Joining [me]$_5$ now is [law professor Rick Pildes, a consultant to [NBC News]$_7$]$_6$. Could [a decision from [the U.S. Supreme Court]$_8$]$_5$ settle [this case]$_2$ once and for [all]$_9$? <TURN> At [this stage]$_{10}$, [any decision from [the U.S. Supreme Court]$_{11}$]$_5$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{13}$]$_{12}$. Indeed, [the issue]$_{14}$ is so narrow now before [the Supreme Court]$_{15}$ that whichever way [the court]$_{16}$ rules, [it]$_2$ will likely have [only the most marginal impact on what's going on in [Florida]$_{18}$]$_{17}$. Even if [the Bush campaign]$_{19}$ prevails before [the Supreme Court]$_{20}$, [it]$_2$ simply means [we]$_{21}$ will move more quickly into [the contest phase of [the litigation]$_{23}$ or [the next stage of [the litigation]$_{23}$]$_{10}$]$_{22}$. <TURN> But [you]$_{24}$ believe [the fact that [the U.S. Supreme Court]$_{25}$ just decided to hear [this case]$_{26}$]$_{19}$ is [a partial victory for [both Bush and Gore]$_{28}$]$_{27}$. <TURN> [It]$_{29}$ is [a partial victory for [both sides]$_{30}$]$_{27}$. For [the last two weeks]$_{31}$, [the central constitutional argument [the Bush campaign]$_{33}$ has been making to [the federal courts]$_{34}$]$_{32}$ is, stop [these manual recounts]$_{35}$ now, [they]$_{36}$ violate [the Constitution]$_{37}$. [The U.S. Supreme Court]$_{29}$ refused to hear [that part of [the case]$_2$]$_{38}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{40}$ is not [the proper time for [federal court intervention]$_{42}$]$_{31}$]$_{41}$. So in [that sense]$_{43}$, [a victory for [the Gore campaign]$_{45}$]$_{44}$. For [the Bush campaign]$_{46}$, [a victory in [the willingness of [the Supreme Court]$_{49}$ to play [some role in overseeing [the Florida system and [the Florida judicial decision making process]$_{52}$]$_{51}$]$_{50}$]$_{48}$]$_{47}$. <TURN> Whatever [the Supreme Court]$_{53}$ decides [this time]$_{41}$, [you]$_{24}$ say [this case]$_2$ could come back before [the U.S. Supreme Court]$_{54}$ again? <TURN> [John]$_{55}$, if [the Supreme Court of [the United States]$_1$]$_{56}$ is to play [a final and decisive role in [this dispute]$_{13}$]$_{57}$, [that role]$_{58}$ is going to come at [the end of [the Florida judicial process]$_{52}$]$_{59}$, not at [this stage]$_{60}$. <TURN> [Law professor Rick Pildes]$_{61}$, thank [you]$_{24}$.

7. **ENCAIX FEBLE**
[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_2$ seems. Joining [me]$_5$ now is [law professor Rick Pildes, a consultant to [NBC News]$_7$]$_6$. Could [a decision from [the U.S. Supreme Court]$_0$]$_5$ settle [this case]$_2$ once and for [all]$_8$? <TURN> At [this stage]$_2$, [any decision from [the U.S. Supreme Court]$_0$]$_0$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{10}$]$_9$. Indeed, [the issue]$_{11}$ is so narrow now before [the Supreme Court]$_0$ that whichever way [the court]$_0$ rules, [it]$_0$ will likely have [only the most marginal impact on what's going on in [Florida]$_{13}$]$_{12}$. Even if [the Bush campaign]$_0$ prevails before [the Supreme Court]$_0$, [it]$_0$ simply means [we]$_{14}$ will move more quickly into [the contest phase of [the litigation]$_{16}$ or [the

next stage of [the litigation]$_{16}$]$_2$]$_{15}$. <TURN> But [you]$_5$ believe [the fact that [the U.S. Supreme Court]$_0$ just decided to hear [this case]$_0$]$_0$ is [a partial victory for [both Bush and Gore]$_{18}$]$_{17}$. <TURN> [It]$_0$ is [a partial victory for [both sides]$_{19}$]$_{17}$. For [the last two weeks]$_{20}$, [the central constitutional argument [the Bush campaign]$_0$ has been making to [the federal courts]$_{22}$]$_{21}$ is, stop [these manual recounts]$_{23}$ now, [they]$_2$ violate [the Constitution]$_{24}$. [The U.S. Supreme Court]$_0$ refused to hear [that part of [the case]$_0$]$_{25}$, agreeing with [all the other federal judges who have unanimously held that [this]$_0$ is not [the proper time for [federal court intervention]$_{27}$]$_0$]$_{26}$. So in [that sense]$_{28}$, [a victory for [the Gore campaign]$_0$]$_{29}$. For [the Bush campaign]$_0$, [a victory in [the willingness of [the Supreme Court]$_0$ to play [some role in overseeing [the Florida system and [the Florida judicial decision making process]$_{33}$]$_{32}$]$_{31}$]$_{30}$]$_{17}$. <TURN> Whatever [the Supreme Court]$_0$ decides [this time]$_0$, [you]$_5$ say [this case]$_0$ could come back before [the U.S. Supreme Court]$_0$ again? <TURN> [John]$_{34}$, if [the Supreme Court of [the United States]$_1$]$_0$ is to play [a final and decisive role in [this dispute]$_{10}$]$_{35}$, [that role]$_0$ is going to come at [the end of [the Florida judicial process]$_{33}$]$_{36}$, not at [this stage]$_0$. <TURN> [Law professor Rick Pildes]$_5$, thank [you]$_0$.

8. **SemEval-2010 RELAXCOR** (SAPENA *et al.*, 2010)
[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_4$ seems. Joining [me]$_5$ now is [law professor Rick Pildes, a consultant to [NBC News]$_7$]$_6$. Could [a decision from [the U.S. Supreme Court]$_9$]$_8$ settle [this case]$_{10}$ once and for [all]$_{11}$? <TURN> At [this stage]$_{12}$, [any decision from [the U.S. Supreme Court]$_{14}$]$_{13}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{16}$]$_{15}$. Indeed, [the issue]$_{17}$ is so narrow now before [the Supreme Court]$_{18}$ that whichever way [the court]$_0$ rules, [it]$_{17}$ will likely have [only the most marginal impact on what's going on in [Florida]$_{20}$]$_{19}$. Even if [the [Bush]$_{22}$ campaign]$_{21}$ prevails before [the Supreme Court]$_{23}$, [it]$_{17}$ simply means [we]$_{24}$ will move more quickly into [the contest phase of [the litigation]$_{26}$ or [the next stage of [the litigation]$_{26}$]$_{27}$]$_{25}$. <TURN> But [you]$_{28}$ believe [the fact that [the U.S. Supreme Court]$_{30}$ just decided to hear [this case]$_{31}$]$_{29}$ is [a partial victory for [both [Bush]$_{34}$ and [Gore]$_{35}$]$_{33}$]$_{32}$. <TURN> [It]$_{29}$ is [a partial victory for [both sides]$_{37}$]$_{36}$. For [the last two weeks]$_{38}$, [the central constitutional argument [the [Bush]$_{40}$ campaign]$_{21}$ has been making to [the federal courts]$_{41}$]$_{39}$ is, stop [these manual recounts]$_{42}$ now, [they]$_{43}$ violate [the Constitution]$_{44}$. [The U.S. Supreme Court]$_{45}$ refused to hear [that part of [the case]$_{47}$]$_{46}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{49}$ is not [the proper time for [federal court intervention]$_{51}$]$_{50}$]$_{48}$. So in [that sense]$_{52}$, [a victory for [the [Gore]$_{54}$ campaign]$_{21}$]$_{53}$. For [the [Bush]$_{55}$ campaign]$_{21}$, [a victory in [the willingness of [the Supreme Court]$_{58}$ to play [some role in overseeing [the [Florida]$_{61}$ system and [the [Florida]$_{63}$ judicial decision making

process]$_{62}$]$_{60}$]$_{59}$]$_{57}$]$_{56}$. <TURN> Whatever [the Supreme Court]$_{64}$ decides [this time]$_{50}$, [you]$_{65}$ say [this case]$_{66}$ could come back before [the U.S. Supreme Court]$_{67}$ again? <TURN> [John]$_{68}$, if [the Supreme Court of [the United States]$_{70}$]$_{69}$ is to play [a final and decisive role in [this dispute]$_{72}$]$_{71}$, [that role]$_{73}$ is going to come at [the end of [the [Florida]$_{76}$ judicial process]$_{75}$]$_{74}$, not at [this stage]$_{77}$. <TURN> [Law professor Rick Pildes]$_{78}$, thank [you]$_{79}$.

9. **SemEval-2010 SUCRE** (KOBDANI i SCHÜTZE, 2010)
[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_4$ seems. Joining [me]$_4$ now is [law professor Rick Pildes, a consultant to [NBC News]$_5$]$_4$. Could [a decision from [the U.S. Supreme Court]$_0$]$_6$ settle [this case]$_2$ once and for [all]$_7$? <TURN> At [this stage]$_8$, [any decision from [the U.S. Supreme Court]$_0$]$_9$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{11}$]$_{10}$. Indeed, [the issue]$_0$ is so narrow now before [the Supreme Court]$_0$ that whichever way [the court]$_0$ rules, [it]$_0$ will likely have [only the most marginal impact on what's going on in [Florida]$_{13}$]$_{12}$. Even if [the [Bush]$_{14}$ campaign]$_0$ prevails before [the Supreme Court]$_0$, [it]$_0$ simply means [we]$_{15}$ will move more quickly into [the contest phase of [the litigation]$_{17}$ or [the next stage of [the litigation]$_{17}$]$_8$]$_{16}$. <TURN> But [you]$_0$ believe [the fact that [the U.S. Supreme Court]$_0$ just decided to hear [this case]$_2$]$_{18}$ is [a partial victory for [both [Bush]$_{14}$ and [Gore]$_{20}$]$_{14}$]$_{19}$. <TURN> [It]$_0$ is [a partial victory for [both sides]$_{22}$]$_{21}$. For [the last two weeks]$_{23}$, [the central constitutional argument [the [Bush]$_{14}$ campaign]$_0$ has been making to [the federal courts]$_{25}$]$_{24}$ is, stop [these manual recounts]$_{26}$ now, [they]$_{27}$ violate [the Constitution]$_{28}$. [The U.S. Supreme Court]$_0$ refused to hear [that part of [the case]$_2$]$_{29}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{31}$ is not [the proper time for [federal court intervention]$_{33}$]$_{32}$]$_{30}$. So in [that sense]$_{34}$, [a victory for [the [Gore]$_{20}$ campaign]$_0$]$_{21}$. For [the [Bush]$_{14}$ campaign]$_0$, [a victory in [the willingness of [the Supreme Court]$_0$ to play [some role in overseeing [the [Florida]$_{13}$ system and [the [Florida]$_{13}$ judicial decision making process]$_{38}$]$_{37}$]$_{36}$]$_{35}$]$_{21}$. <TURN> Whatever [the Supreme Court]$_0$ decides [this time]$_{32}$, [you]$_0$ say [this case]$_2$ could come back before [the U.S. Supreme Court]$_0$ again? <TURN> [John]$_{39}$, if [the Supreme Court of [the United States]$_{40}$]$_0$ is to play [a final and decisive role in [this dispute]$_{11}$]$_{36}$, [that role]$_{36}$ is going to come at [the end of [the [Florida]$_{13}$ judicial process]$_{38}$]$_{41}$, not at [this stage]$_8$. <TURN> [Law professor Rick Pildes]$_4$, thank [you]$_4$.

10. **SemEval-2010 TANL-1** (ATTARDI *et al.*, 2010)
[[The nation's]$_1$ highest court]$_0$ will take up [the case]$_2$ [next week]$_3$. [That development]$_4$ may not be as significant as [it]$_5$ seems. Joining [me]$_6$ now is

[law professor [Rick Pildes]$_7$, a consultant to [NBC News]$_8$]$_7$. Could [a decision from [the [U.S. Supreme Court]$_{11}$]$_{10}$]$_9$ settle [this case]$_{11}$ once and for [all]$_{12}$? <TURN> At [this stage]$_{13}$, [any decision from [the [U.S. Supreme Court]$_{11}$]$_{10}$]$_{14}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{16}$]$_{15}$. Indeed, [the issue]$_{17}$ is so narrow now before [the [Supreme Court]$_{11}$]$_{10}$ that whichever way [the court]$_{18}$ rules, [it]$_{19}$ will likely have [only the most marginal impact on what's going on in [Florida]$_{21}$]$_{20}$. Even if [the [Bush]$_{23}$ campaign]$_{22}$ prevails before [the [Supreme Court]$_{11}$]$_{10}$, [it]$_{24}$ simply means [we]$_{25}$ will move more quickly into [the contest phase of [the litigation]$_{27}$ or [the next stage of [the litigation]$_{29}$]$_{28}$]$_{26}$. <TURN> But [you]$_{30}$ believe [the fact that [the [U.S. Supreme Court]$_{11}$]$_{10}$ just decided to hear [this case]$_{32}$]$_{31}$ is [a partial victory for [both [Bush]$_{35}$ and [Gore]$_{36}$]$_{34}$]$_{33}$. <TURN> [It]$_{37}$ is [a partial victory for [both sides]$_{39}$]$_{38}$. For [the last two weeks]$_{40}$, [the central constitutional argument [the [Bush]$_{35}$ campaign]$_{42}$ has been making to [the federal courts]$_{43}$]$_{41}$ is, stop [these manual recounts]$_{44}$ now, [they]$_{45}$ violate [the Constitution]$_{46}$. [The [U.S. Supreme Court]$_{11}$]$_{10}$ refused to hear [that part of [the case]$_{48}$]$_{47}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{50}$ is not [the proper time for [federal court intervention]$_{52}$]$_{51}$]$_{49}$. So in [that sense]$_{53}$, [a victory for [the [Gore]$_{36}$ campaign]$_{55}$]$_{54}$. For [the [Bush]$_{35}$ campaign]$_{56}$, [a victory in [the willingness of [the [Supreme Court]$_{11}$]$_{59}$ to play [some role in overseeing [the [Florida]$_{62}$ system and [the [Florida]$_{62}$ judicial decision making process]$_{63}$]$_{61}$]$_{60}$]$_{58}$]$_{57}$. <TURN> Whatever [the [Supreme Court]$_{11}$]$_{59}$ decides [this time]$_{64}$, [you]$_{65}$ say [this case]$_{66}$ could come back before [the [U.S. Supreme Court]$_{68}$]$_{67}$ again? <TURN> [John]$_{69}$, if [the [Supreme Court]$_{11}$ of [the [United States]$_{71}$]$_{70}$]$_{67}$ is to play [a final and decisive role in [this dispute]$_{73}$]$_{72}$, [that role]$_{74}$ is going to come at [the end of [the [Florida]$_{77}$ judicial process]$_{76}$]$_{75}$, not at [this stage]$_{78}$. <TURN> [Law professor [Rick Pildes]$_7$]$_{79}$, thank [you]$_{80}$.

11. **SemEval-2010 UBIU** (ZHEKOVA i KÜBLER, 2010)
[The nation's highest court]$_0$ will take up [the case]$_1$ [next week]$_2$. [That development]$_3$ may not be as significant as [it]$_3$ seems. Joining [me]$_0$ now is [law professor Rick Pildes, a consultant to NBC News]$_4$. Could [a decision from [the U.S. Supreme Court]$_5$]$_4$ settle [this case]$_5$ once and for [all]$_6$? <TURN> At [this stage]$_5$, [any decision from [the U.S. Supreme Court]$_5$]$_7$ is almost certainly not going to provide [a final resolution of [this election dispute]$_9$]$_8$. Indeed, [the issue]$_{10}$ is so narrow now before [the Supreme Court]$_5$ that [whichever way the court]$_{11}$ rules, [it]$_{11}$ will likely have [only the most marginal impact on what's going on in Florida]$_{12}$. Even if [the [Bush]$_{14}$ campaign]$_{13}$ prevails before [the Supreme Court]$_{15}$, [it]$_{11}$ simply means [we]$_5$ will move more quickly into [the contest phase of [the litigation]$_{17}$ or [the next stage of [the litigation]$_{19}$]$_{18}$]$_{16}$. <TURN> But [you]$_{20}$ believe [the fact that [the U.S. Supreme Court]$_{15}$ just decided to he-

ar [this case]$_{22}$]$_{21}$ is [a partial victory for [both [Bush]$_{14}$ and [Gore]$_{25}$]$_{24}$]$_{23}$. <TURN> [It]$_{26}$ is [a partial victory for [both sides]$_{28}$]$_{27}$. For [the last two weeks]$_{29}$, [the central constitutional argument [the [Bush]$_{14}$ campaign]$_{31}$ has been making to [the federal courts]$_{32}$]$_{30}$ is, stop [these manual recounts]$_{33}$ now, [they]$_{20}$ violate [the Constitution]$_{34}$. [The U.S. Supreme Court]$_{35}$ refused to hear [that part of [the case]$_{36}$]$_{35}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{38}$ is not [the proper time for [federal court intervention]$_{40}$]$_{39}$]$_{37}$. So in [that sense]$_{41}$, [a victory for [the [Gore]$_{43}$ campaign]$_{41}$]$_{42}$. For [the [Bush]$_{45}$ campaign]$_{44}$, [a victory in [the willingness of [the Supreme Court]$_{48}$ to play [some role in overseeing [the [Florida]$_{51}$ system and [the [Florida]$_{53}$ judicial decision making process]$_{52}$]$_{50}$]$_{49}$]$_{47}$]$_{46}$. <TURN> Whatever [the Supreme Court]$_{48}$ decides [this time]$_{49}$, [you]$_{48}$ say [this case]$_{50}$ could come back before [the U.S. Supreme Court]$_{51}$ again? <TURN> [John]$_{52}$, if [the Supreme Court of [the United States]$_{54}$]$_{53}$ is to play [a final and decisive role in this dispute]$_{55}$, [that role]$_{56}$ is going to come at [the end of the [Florida]$_{58}$ judicial process]$_{57}$, not at [this stage]$_{59}$. <TURN> [Law professor Rick Pildes]$_{60}$, thank [you]$_{48}$.

12. **SemEval-2010 CORRY-C** (URYUPINA, 2010)
[[The nation's]$_{1}$ highest court]$_{0}$ will take up [the case]$_{2}$ [next week]$_{3}$. [That development]$_{4}$ may not be as significant as [it]$_{5}$ seems. Joining [me]$_{4}$ now is [law professor Rick Pildes, a consultant to [NBC News]$_{7}$]$_{6}$. Could [a decision from [the U.S. Supreme Court]$_{9}$]$_{8}$ settle [this case]$_{2}$ once and for [all]$_{10}$? <TURN> At [this stage]$_{11}$, [any decision from [the U.S. Supreme Court]$_{9}$]$_{8}$ is almost certainly not going to provide [a final resolution of [this election dispute]$_{13}$]$_{12}$. Indeed, [the issue]$_{14}$ is so narrow now before [the Supreme Court]$_{9}$ that whichever way [the court]$_{15}$ rules, [it]$_{14}$ will likely have [only the most marginal impact on what's going on in [Florida]$_{17}$]$_{16}$. Even if [the [Bush]$_{19}$ campaign]$_{18}$ prevails before [the Supreme Court]$_{9}$, [it]$_{14}$ simply means [we]$_{20}$ will move more quickly into [the contest phase of [the litigation]$_{22}$ or [the next stage of [the litigation]$_{22}$]$_{23}$]$_{21}$. <TURN> But [you]$_{24}$ believe [the fact that [the U.S. Supreme Court]$_{9}$ just decided to hear [this case]$_{2}$]$_{25}$ is [a partial victory for [both [Bush]$_{19}$ and [Gore]$_{28}$]$_{27}$]$_{26}$. <TURN> [It]$_{14}$ is [a partial victory for [both sides]$_{30}$]$_{29}$. For [the last two weeks]$_{31}$, [the central constitutional argument [the [Bush]$_{19}$ campaign]$_{18}$ has been making to [the federal courts]$_{33}$]$_{32}$ is, stop [these manual recounts]$_{34}$ now, [they]$_{31}$ violate [the Constitution]$_{35}$. [The U.S. Supreme Court]$_{9}$ refused to hear [that part of [the case]$_{2}$]$_{36}$, agreeing with [all the other federal judges who have unanimously held that [this]$_{38}$ is not [the proper time for [federal court intervention]$_{40}$]$_{39}$]$_{37}$. So in [that sense]$_{41}$, [a victory for [the [Gore]$_{28}$ campaign]$_{43}$]$_{42}$. For [the [Bush]$_{41}$ campaign]$_{19}$, [a victory in [the willingness of [the Supreme Court]$_{9}$ to play [some role in overseeing [the [Florida]$_{17}$ system and [the [Florida]$_{47}$ judicial decision making process]$_{17}$]$_{46}$]$_{45}$]$_{44}$]$_{18}$. <TURN> Whatever [the Supreme Court]$_{9}$ decides

[this time]$_{48}$, [you]$_{24}$ say [this case]$_{2}$ could come back before [the U.S. Supreme Court]$_{9}$ again? <TURN> [John]$_{49}$, if [the Supreme Court of [the United States]$_{50}$]$_{9}$ is to play [a final and decisive role in [this dispute]$_{52}$]$_{51}$, [that role]$_{45}$ is going to come at [the end of [the [Florida]$_{17}$ judicial process]$_{54}$]$_{53}$, not at [this stage]$_{11}$. <TURN> [Law professor Rick Pildes]$_{6}$, thank [you]$_{24}$.

13. **SemEval-2010 BART** (BROSCHEIT *et al.*, 2010)
[The nation's]$_{0}$ highest court will take up [the case]$_{1}$ [next week]$_{2}$. [That development]$_{3}$ may not be as significant as [it]$_{4}$ seems. Joining [me]$_{5}$ now is [[law]$_{7}$ professor Rick Pildes]$_{6}$, [a consultant]$_{8}$ to [NBC News]$_{9}$. Could [a decision]$_{10}$ from the U.S. [Supreme Court]$_{11}$ settle [this case]$_{1}$ once and for all? <TURN> At [this stage]$_{12}$, [any decision]$_{10}$ from the [U.S. Supreme Court]$_{11}$ is almost certainly not going to provide [a final resolution]$_{13}$ of [this [election]$_{15}$ dispute]$_{14}$. Indeed, [the issue]$_{16}$ is so narrow now before the [Supreme Court]$_{17}$ that whichever way [the court]$_{18}$ rules, [it]$_{4}$ will likely have [only the most marginal impact]$_{19}$ on what's going on in [Florida]$_{20}$. Even if [the [Bush]$_{22}$ campaign]$_{21}$ prevails before the [Supreme Court]$_{17}$, [it]$_{4}$ simply means [we]$_{23}$ will move more quickly into [the [contest]$_{25}$ phase]$_{24}$ of [the litigation]$_{26}$ or [the next stage of [the litigation]$_{26}$]$_{12}$. <TURN> But [you]$_{27}$ believe the fact that the [U.S. Supreme Court]$_{11}$ just decided to hear [this case]$_{1}$ is [a partial victory]$_{28}$ for [both [Bush]$_{22}$ and [Gore]$_{30}$]$_{29}$. <TURN> [It]$_{11}$ is [a partial victory]$_{28}$ for [both sides]$_{31}$. For [the last two weeks]$_{32}$, [the central constitutional argument]$_{33}$ [the [Bush]$_{22}$ campaign]$_{21}$ has been making to [the federal courts]$_{34}$ is, stop [these manual recounts]$_{35}$ now, [they]$_{36}$ violate [the Constitution]$_{37}$. The [U.S. Supreme Court]$_{11}$ refused to hear that [part]$_{38}$ of [the case]$_{1}$, agreeing with [all the other federal judges]$_{39}$ who have unanimously held that [this]$_{40}$ is not [the proper time]$_{41}$ for [federal [court]$_{18}$ intervention]$_{42}$. So in [that sense]$_{43}$, [a victory]$_{44}$ for [the [Gore]$_{30}$ campaign]$_{21}$. For [the [Bush]$_{22}$ campaign]$_{21}$, [a victory]$_{44}$ in [the willingness]$_{45}$ of the [Supreme Court]$_{17}$ to play [some role]$_{46}$ in overseeing [the [Florida]$_{20}$ system]$_{47}$ and [the [[Florida]$_{20}$ judicial decision]$_{49}$ making process]$_{48}$. <TURN> Whatever the [Supreme Court]$_{17}$ decides [this time]$_{41}$, [you]$_{27}$ say [this case]$_{1}$ could come back before the [U.S. Supreme Court]$_{11}$ again? <TURN> [John]$_{50}$, if the [Supreme Court]$_{17}$ of [the United States]$_{51}$ is to play [a final and decisive role]$_{52}$ in [this dispute]$_{14}$, [that role]$_{46}$ is going to come at [the end]$_{53}$ of [the [Florida]$_{20}$ judicial process]$_{48}$, not at [this stage]$_{12}$. <TURN> [[Law]$_{7}$ professor Rick Pildes]$_{6}$, thank [you]$_{27}$.

A.2 Fitxer d'OntoNotes voa_0207

Cuban leader Fidel Castro is setting up a lavish extravaganza on the island nation to welcome the new millennium, one year late for much of the rest of the world. Many experts contend most of the world was at least technically wrong by bringing in the new millennium with massive celebrations last year. These experts point out that the Gregorian calendar started in 1 AD and therefore, centuries' millennia start with a one, not a zero. They say this makes 2001 the first year of the third millennium. For those observing the start of 2001 as a true dawn of the twenty-first century, the parties and fireworks are fewer and less elaborate than the 2000 celebrations. In Cuba though, where President Castro had his country sit out last year's revelry, they'll be making up for it as major festivities are set.

1. **ESTÀNDARD DE REFERÈNCIA**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_4$ late for [much of [the rest of [the world]$_7$]$_6$]$_5$. [Many experts]$_8$ contend [most of [the world]$_7$]$_9$ was at least technically wrong by bringing in [the new millennium]$_3$ with [massive celebrations]$_{10}$ [last year]$_{11}$. [These experts]$_8$ point out that [the Gregorian calendar]$_{12}$ started in [1 AD]$_{13}$ and therefore, [[centuries']$_{15}$ millennia]$_{14}$ start with [a one, not [a zero]$_{17}$]$_{16}$. [They]$_8$ say [this]$_{18}$ makes [2001]$_{19}$ [the first year of [the third millennium]$_3$]$_{20}$. For [those observing [the start of [2001]$_{19}$]$_{22}$ as [a true dawn of [the twenty-first century]$_{24}$]$_{23}$]$_{21}$, [the parties and fireworks]$_{25}$ are fewer and less elaborate than [the 2000 celebrations]$_{26}$. In [Cuba]$_2$ though, where [President Castro]$_0$ had [[his]$_0$ country]$_2$ sit out [[last year's]$_{11}$ revelry]$_{26}$, [they]$_2$'ll be making up for [it]$_{27}$ as [major festivities]$_{28}$ are set.

2. **BASELINE: TOTES UNÀRIES**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_4$ late for [much of [the rest of [the world]$_7$]$_6$]$_5$. [Many experts]$_8$ contend [most of [the world]$_{10}$]$_9$ was at least technically wrong by bringing in [the new millennium]$_{11}$ with [massive celebrations]$_{12}$ [last year]$_{13}$. [These experts]$_{14}$ point out that [the Gregorian calendar]$_{15}$ started in [1 AD]$_{16}$ and therefore, [[centuries']$_{18}$ millennia]$_{17}$ start with [a one, not [a zero]$_{20}$]$_{19}$. [They]$_{21}$ say [this]$_{22}$ makes [2001]$_{23}$ [the first year of [the third millennium]$_{25}$]$_{24}$. For [those observing [the start of [2001]$_{28}$]$_{27}$ as [a true dawn of [the twenty-first century]$_{30}$]$_{29}$]$_{26}$, [the parties and fireworks]$_{31}$ are fewer and less elaborate than [the 2000 celebrations]$_{32}$. In [Cuba]$_{33}$ though, where [President Castro]$_{34}$ had [[his]$_{36}$ country]$_{35}$ sit out [[last year's]$_{38}$ revelry]$_{37}$, [they]$_{39}$'ll be making up for [it]$_{40}$ as [major festivities]$_{41}$ are set.

3. **BASELINE: MATEIX NUCLI**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_4$ late for [much of [the rest of [the world]$_7$]$_6$]$_5$. [Many experts]$_8$ contend [most of [the world]$_7$]$_9$ was at least technically wrong by bringing in [the new millennium]$_3$ with [massive celebrations]$_{10}$ [last year]$_4$. [These experts]$_8$ point out that [the Gregorian calendar]$_{11}$ started in [1 AD]$_{12}$ and therefore, [[centuries']$_{14}$ millennia]$_{13}$ start with [a one, not [a zero]$_{16}$]$_{15}$. [They]$_{17}$ say [this]$_{18}$ makes [2001]$_{19}$ [the first year of [the third millennium]$_3$]$_4$. For [those observing [the start of [2001]$_{19}$]$_{21}$ as [a true dawn of [the twenty-first century]$_{23}$]$_{22}$]$_{20}$, [the parties and fireworks]$_{24}$ are fewer and less elaborate than [the 2000 celebrations]$_{10}$. In [Cuba]$_{25}$ though, where [President Castro]$_0$ had [[his]$_{27}$ country]$_{26}$ sit out [[last year's]$_4$ revelry]$_{28}$, [they]$_{29}$'ll be making up for [it]$_{30}$ as [major festivities]$_{31}$ are set.

4. **ENCAIX FORT**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_3$ late for [much of [the rest of [the world]$_6$]$_5$]$_4$. [Many experts]$_7$ contend [most of [the world]$_6$]$_8$ was at least technically wrong by bringing in [the new millennium]$_3$ with [massive celebrations]$_9$ [last year]$_3$. [These experts]$_{10}$ point out that [the Gregorian calendar]$_{11}$ started in [1 AD]$_{12}$ and therefore, [[centuries']$_{14}$ millennia]$_{13}$ start with [a one, not [a zero]$_{16}$]$_{15}$. [They]$_{10}$ say [this]$_{17}$ makes [2001]$_{18}$ [the first year of [the third millennium]$_3$]$_3$. For [those observing [the start of [2001]$_{18}$]$_{20}$ as [a true dawn of [the twenty-first century]$_{13}$]$_{21}$]$_{19}$, [the parties and fireworks]$_{22}$ are fewer and less elaborate than [the 2000 celebrations]$_9$. In [Cuba]$_{23}$ though, where [President Castro]$_0$ had [[his]$_0$ country]$_{24}$ sit out [[last year's]$_{26}$ revelry]$_{25}$, [they]$_7$'ll be making up for [it]$_{19}$ as [major festivities]$_{27}$ are set.

5. **ENCAIX SUPERFORT**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_4$ late for [much of [the rest of [the world]$_7$]$_6$]$_5$. [Many experts]$_8$ contend [most of [the world]$_7$]$_9$ was at least technically wrong by bringing in [the new millennium]$_3$ with [massive celebrations]$_{10}$ [last year]$_{11}$. [These experts]$_{12}$ point out that [the Gregorian calendar]$_{13}$ started in [1 AD]$_{14}$ and therefore, [[centuries']$_{16}$ millennia]$_{15}$ start with [a one, not [a zero]$_{18}$]$_{17}$. [They]$_{12}$ say [this]$_{19}$ makes [2001]$_{20}$ [the first year of [the third millennium]$_3$]$_{21}$. For [those observing [the start of [2001]$_{20}$]$_{23}$ as [a true dawn of [the twenty-first century]$_{15}$]$_{24}$]$_{22}$, [the parties and fireworks]$_{25}$ are fewer and less elaborate than [the 2000 celebrations]$_{10}$. In [Cuba]$_{26}$ though, where [President Castro]$_0$ had [[his]$_0$ country]$_{27}$ sit out [[last year's]$_{29}$ revelry]$_{28}$, [they]$_8$'ll be making up for [it]$_{21}$ as [major festivities]$_{30}$ are set.

6. **ENCAIX ÒPTIM**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_3$ late for [much of [the rest of [the world]$_6$]$_5$]$_4$. [Many experts]$_7$ contend [most of [the world]$_6$]$_8$ was at least technically wrong by bringing in [the new millennium]$_3$ with [massive celebrations]$_9$ [last year]$_3$. [These experts]$_{10}$ point out that [the Gregorian calendar]$_{11}$ started in [1 AD]$_{12}$ and therefore, [[centuries']$_{14}$ millennia]$_{13}$ start with [a one, not [a zero]$_{16}$]$_{15}$. [They]$_{10}$ say [this]$_{17}$ makes [2001]$_{18}$ [the first year of [the third millennium]$_3$]$_3$. For [those observing [the start of [2001]$_{18}$]$_{20}$ as [a true dawn of [the twenty-first century]$_{13}$]$_{21}$]$_{19}$, [the parties and fireworks]$_{22}$ are fewer and less elaborate than [the 2000 celebrations]$_9$. In [Cuba]$_{23}$ though, where [President Castro]$_0$ had [[his]$_0$ country]$_{24}$ sit out [[last year's]$_{26}$ revelry]$_{25}$, [they]$_{10}$'ll be making up for [it]$_{19}$ as [major festivities]$_{27}$ are set.

7. **ENCAIX FEBLE**
   [Cuban leader Fidel Castro]$_0$ is setting up [a lavish extravaganza on [the island nation]$_2$]$_1$ to welcome [the new millennium]$_3$, [one year]$_3$ late for [much of [the rest of [the world]$_6$]$_5$]$_4$. [Many experts]$_7$ contend [most of [the world]$_6$]$_8$ was at least technically wrong by bringing in [the new millennium]$_3$ with [massive celebrations]$_9$ [last year]$_3$. [These experts]$_{10}$ point out that [the Gregorian calendar]$_{11}$ started in [1 AD]$_{12}$ and therefore, [[centuries']$_{14}$ millennia]$_{13}$ start with [a one, not [a zero]$_{16}$]$_{15}$. [They]$_{10}$ say [this]$_{17}$ makes [2001]$_{18}$ [the first year of [the third millennium]$_3$]$_3$. For [those observing [the start of [2001]$_{18}$]$_{20}$ as [a true dawn of [the twenty-first century]$_3$]$_{21}$]$_{19}$, [the parties and fireworks]$_{10}$ are fewer and less elaborate than [the 2000 celebrations]$_9$. In [Cuba]$_{22}$ though, where [President Castro]$_0$ had [[his]$_0$ country]$_{23}$ sit out [[last year's]$_3$ revelry]$_{24}$, [they]$_7$'ll be making up for [it]$_3$ as [major festivities]$_{25}$ are set.

APÈNDIX B

Fragments de quasi-identitat

Aquest apèndix inclou el corpus de 60 fragments que es va utilitzar en grups de 20 als tres experiments descrits a l'apartat 8.5. Els fragments estan extrets de tres corpus electrònics –ACE (DODDINGTON *et al.*, 2004), OntoNotes (PRADHAN *et al.*, 2007*a*) i AnCora (RECASENS i MARTÍ, 2010)– així com d'internet, un programa de televisió i conversa real.

En la tasca es demanava als anotadors que classifiquessin en un dels tipus de relació de (quasi-)identitat (apartat 8.4) els parells de SNs entre claudàtors de cada fragment. Havien d'assignar un o més tipus, però com a mínim un, a cada parell de SNs. Les respostes es recullen a l'apèndix C.

B.1 Experiment 1

(1) [Firestone]$_1$ chairman John Lampe, on a telephone conference call with reporters this afternoon . . . I see the concern in people's faces. And they're very apprehensive about purchasing [Firestones]$_2$.

(2) Hoddle does not resign after his opinion about [the disabled]$_1$. The Times had published some declarations of the English manager in which he said that "[the physically and mentally disabled]$_2$ pay for the sins they committed in a previous life."

(3) [A beloved American holiday story]$_1$ comes to the big screen in [a Universal Pictures comic fantasy starring Jim Carey]$_2$. Alan Silverman has a look at the first feature film adaptation of Dr. Seuss's *How the Grinch Stole Christmas* ... [it]$_3$'s the whimsical story of the Grinch ... Director Ron Howard set out to film [the fantasy]$_4$, not as a cartoon, but with actors in costumes and settings in the spirit of [the book]$_5$.

(4) Juan Carlos Ferrero and Francisco Clavet, the two last hopes of Spanish male tennis in [the Australian Open]$_1$, were eliminated today in the third round ... Ferrero had become one of the revelations of [the tournament]$_2$ ... It is his best performance in the [Australian Open, where he had never progressed past the second round]$_3$.

(5) As the sun rises over [Mt. Popo]$_1$ tonight, the only hint of the fire storm inside, whiffs of smoke ... [The fourth largest mountain in North America, nearly 18,000 feet high]$_2$, erupting this week with its most violent outburst in 1,200 years.

(6) [US]$_1$ victims of terrorism have been able to sue foreign governments since 1996. But under legislation passed this month, many victims will actually get their money. The money, at least at first, will come from the US treasury. [The government]$_2$ expects to get it back from frozen Iranian assets held in [this country]$_3$.

(7) It's the whimsical story of the Grinch, a mean spirited hairy green creature who menaces the holiday loving Hus until an innocent child Mary Lu Hu teaches him to find the joy in life ... [Starter Jim Carey]$_1$ says the Grinch is more than just a cold hearted character. [He]$_2$ is the outcast ... [Carey]$_3$ performs covered head to toe in that green-haired costume ... Oh, you will recognize [me]$_4$.

(8) The gigantic international auction house Sotheby's pleaded guilty to price-fixing with Christie's—its only real competition in an industry that does $4 billion in business every year ... [The cartel]$_1$ consisted of [Sotheby's and Christie's]$_2$. [Arch rivals for nearly three centuries, the two auction houses]$_3$ agreed to fix prices on what [they]$_4$ charged the buyers and sellers of high-priced art ... [Sotheby's and Christie's]$_5$ are all about money.

(9) In France, [the president]$_1$ is elected for a term of seven years, while in the United States [he]$_2$ is elected for a term of four years.

(10) Fishermen on this Canadian island province have shared tales of their catch. Lobster in recent years. But not too long ago, [another delicacy

– salmon]$_1$. Oh, yeah, we used to get [salmon]$_2$ in the spring, but we don't see [it]$_3$ anymore. I think [they]$_4$ are pretty well wiped out ... it's important people know if creating supersalmon to feed human appetites could threaten [normal salmon]$_5$.

(11) Montse Aguer claimed that there is an image of [Dalí]$_1$, which is the easiest one: [the provocative Dalí]$_2$, whose most popular works are known.

(12) Juan Antonio Samaranch asked the Australian city to provide [certain information]$_1$ ... President Samaranch sent a letter to Sydney in which he asked for [information]$_2$.

(13) —has in the world, one in the Middle East is all too obvious, and as of is in broadcast tonight, the Clinton administration is not making much progress getting Palestinians and Israelis to lay off each other and talk about it. The other is [North Korea]$_1$ ... We'll get to [Korea]$_2$ in a minute.

(14) On homecoming night [Postville]$_1$ feels like Hometown, USA, but a look around this town of 2,000 shows [it]$_2$'s become a miniature Ellis Island. [This]$_3$ was an all-white, all-christian community that all the sudden was taken over – not taken over, that's a very bad choice of words, but invaded by, perhaps, different groups ... [Postville]$_4$ now has 22 different nationalities ... For those who prefer [the old Postville]$_5$, Mayor John Hyman has a simple answer.

(15) A study of nearly 300 former British professional soccer players finds that [nearly half]$_1$ suffered the chronic joint disease "osteoarthritis" often as early as age 40. Most have the disease in two or more joints ... The Coventry University researchers who report the findings in the British journal of sports medicine say anxiety and depression are common among [those so injured]$_2$.

(16) In many cities, [angry crowds]$_1$ roam the streets, [Jews]$_2$ and Palestinians looking for confrontation. Last night in Tel Aviv, [Jews]$_3$ attacked a restaurant that employs Palestinians "[we]$_4$ want war," [the crowd]$_5$ chanted.

(17) [The trial thrust chief prosecutor Marcia Clark]$_1$ into the spotlight. [Clark]$_2$ graduated from UCLA in 1974, earning her law degree five years later ... Clark gained reputation for her expertise in forensic evidence, handling at least 60 jury trials, 20 involving murder ... the Simpson trial and the jury's findings marked a turning point in the career of [the twice-divorced mother of two]$_3$.

(18) US Energy Secretary, Bill Richardson, has extended [an emergency order]$_1$ to keep electricity flowing to California. [The measure]$_2$ will require Western suppliers to sell power to the State for at least another week.

(19) The rate of increase of [the December 2000 CPI in entire Spain]$_1$ stayed at the 2.9 per cent ... Regarding Catalonia, [the CPI]$_2$ stays at the 3.5 per cent.

(20) We begin tonight with [the huge federal surplus]$_1$. Both Al Gore and Bush have different ideas on how to spend [that extra money]$_2$. The last time presidential candidates had [that luxury]$_3$ was in 1960.

B.2 Experiment 2

(21) [Egypt]$_1$ needs more than 250 million dollars to eliminate the mine camps that are found in different areas of [this country]$_2$.

(22) Half of children under the age of 5 get [the flu]$_1$. While unvaccinated kids bring [it]$_2$ home and infect brothers and sisters, a vaccinated child helps reduce the risk by 80%.

(23) That's according to [a new study from the secret service on school violence]$_1$. [It]$_2$ shows that attackers, like the two who killed 13 people at Columbine High School last year in Colorado, come from a variety of family and ethnic backgrounds. Academic performance ranged from excellent to failure ... [It]$_3$'s really a fact-based report. And with [these facts]$_4$, a school can move out and actually do prevention.

(24) Patricia Ferreira makes progress making thriller films with [her second feature film, *The Impatient Alchemist*, presented yesterday in the competition section of the Spanish Film Festival]$_1$. [The film, based on [the novel of the same title by Lorenzo Silva]$_2$]$_3$, is a thriller ... [It]$_4$ has different readings, an original plot and the portrait of a society, which is ours.

(25) [An International team]$_1$ is developing a vaccine against Alzheimer's disease and [they]$_2$ are trying it out on a new and improved mouse model of the onus ... [Scientists working on a vaccine against Alzheimer's]$_3$ give a progress report this week in the journal Nature.

(26) The Barcelona Chamber of Commerce has marked [the Catalan GDP growth]$_1$ during last year in 3.7 per cent ... Regarding the growth of the

economy during last year's last three months, [the GDP growth figures]$_2$ reached 3.9 per cent, three tenths over [that obtained in the previous months]$_3$.

(27) (*Halle Berry speaking*) I am in the supermarket and I was just on the cover of Bazaar magazine. At an early age my daughter would recognize [me]$_1$ in the photo... so I've got on my sunglasses and I'm in the market, I'm putting on my groceries ... and she's over my shoulder and I hear her say ["Mama, mama"]$_2$, and I knew "Oh, she saw that cover, that's cute." And this woman behind her was sort of cooing with her, and I heard the woman say "Oh, no, honey, [that]$_3$'s not your mama, that's Halle Berry." "Mama, mama." And the lady sort of got like indignant about it: "No, honey, that's not your mama, that's Halle Berry." And I couldn't take it any longer: "No, [I]$_4$ am her mother and [I]$_5$ am Halle Berry, and she knows what she's talking about."

(28) The Catalan Corporation of Radio and Television joined today [the Year of Dalí 2004]$_1$ as a participating institution. The general director of the Catalan Corporation of Radio and Television stated that talking about Dalí in [2004]$_2$ does not require much effort.

(29) The trade union representing performers and the agents of [Hollywood]$_1$ continue their conversations ... They ask for a 5% increase and [the studios]$_2$ offer a 3.55%.

(30) We're joined by NBC news correspondent Campbell Brown Ho who's traveling with [the Bush effort]$_1$ ... Bush's central message on this bus trip across central Florida today was to his diehard supporters telling them go out, tell your friends still on the fence why they need to vote for me. And it's a message [the campaign]$_2$ hopes [it]$_3$ was able to convey today. Because while Florida is a must win, [they]$_4$ also cannot ignore the other battleground states.

(31) The strategy has been a popular one for [McDonalds]$_1$, as a sample poll of lunchtime customers outside a restaurant in South Delhi shows ... Here, you know, it's like it's American and as well as Indian taste. It's a very wise move on for them because if they would have [only just original McDonalds]$_2$, I don't think they would have done so great.

(32) The Prime Minister, José María Aznar, said today that the twenty-five years of reign of Juan Carlos I "have been successful and extraordinarily important for [Spain]$_1$" ... According to Aznar, Parliamentary Monarchy "is not only the expression of [the modern Spain]$_2$, but it is also a symbol

of stability and permanence."

(33) [Two populations with different backgrounds]$_1$ work as specialist doctors. One, MIR, which follows a government-regulated education ... The other one, the turkey oak ... followed heterogeneous education and training formulas ... The comparative study of [two cohorts with these characteristics]$_2$ was the object of my PhD thesis.

(34) It's acquiring [more pressure]$_1$. And eventually [this pressure]$_2$ will be released in the – in the future days.

(35) Juan Antonio Samaranch did not order starting an investigation about [Sydney-2000]$_1$, but asked [the Australian city]$_2$ to provide certain information.

(36) The figure of Dalí was born in [a Catalan cultural context]$_1$. We are simply remembering that Dalí was born from [the Catalan cultural context]$_2$.

(37) Nader condemns corporations, drug companies, pesticide manufacturers, banks, landlords, the media. [His supporters]$_1$ say [they]$_2$ don't care that he has no chance to become President.

(38) Tony Blair lamented the declarations of [the English manager]$_1$ and he showed his preference for him to abandon [his position]$_2$.

(39) If [the United States]$_1$ has officially restored diplomatic relations with Yugoslavia, [President Clinton]$_2$ announced the move during his visit to Vietnam ... [The White House]$_3$ said [the United States]$_4$ will provide 45 million dollars in food aid to Yugoslavia.

(40) The ex Real Madrid player is the only change in the list, comprised of [18 soccer players]$_1$. Eto'o said that [the team]$_2$ should not be too confident because of the result of the first leg of Copa del Rey.

B.3 Experiment 3

(41) Five years ago [today]$_1$, the O.J. Simpson trial ended with an acquittal ... On [this day in 1995]$_2$, O.J. Simpson was acquitted of the 1994 murders of his ex-wife Nicole and her friend Ron Goldman.

(42) [The Denver Broncos]$_1$ assure their Super Bowl title. [Denver]$_2$ was led by a great John Elway.

(43) Meanwhile, at the Sun Ball in El Paso Texas, the University of Wisconsin Badgers held off [the University of California at Los Angeles]$_1$ 21-20. The Badger's coach Barry Averett says that [his seniors]$_2$ showed leadership in making their last game one of their best. [We]$_3$ were soft after that first drop. Sometimes when it comes too easy you can get soft but I liked the way [they]$_4$ responded.

(44) [*When we see each other*]$_1$ is the title of the last record of the band Bars. [It]$_2$ contains songs from their six records.

(45) But only two miles away, [Atlantic salmon]$_1$ are thriving, and that's an understatement. [These experimental salmon]$_2$ are on the cutting edge of the debate over genetically engineered food ... it's important people know if creating [supersalmon to feed human appetites]$_3$ could threaten normal salmon. We have shown [they]$_4$ have a tremendous potential to upset the balance of nature.

(46) The strategy has been a popular one for [McDonalds]$_1$, as a sample poll of lunchtime customers outside a restaurant in South Delhi shows ... Here, you know, it's like it's American and as well as Indian taste. It's a very wise move on for [them]$_2$.

(47) The US government is warning American citizens living and traveling abroad to be on alert as [violence]$_1$ continues in the Mideast. [The confrontations]$_2$ are casting a shadow over Mideast peace talks in Paris ... He wants the Israelis to end [the fighting]$_3$.

(48) [Britain's Millennium Dome]$_1$ will close down this coming Monday after a year of mishaps ... Problems riddled [the Dome]$_2$ even before its grand opening last New Year's Eve ... Dome officials had to seek an additional 265 million dollars to complete [the structure]$_3$.

(49) The director of the Catalan Corporation of Radio and Television stated that talking about [Dalí]$_1$ in 2004 does not require much effort ... it is the moment when we must define [the figure of Dalí]$_2$.

(50) Yugoslav opposition leaders criticized [the United States]$_1$ and Russia today as a strike against President Slobodan Milosevic gained momentum across the country ... Kostunica accused the Russian government of indecision and said [Washington]$_2$ was indirectly helping Milosevic's cause.

(51) Textbooks provide students with an equilibrated view of [the history of Spain]$_1$. A report from the Real Academy of History released this week accused the autonomous communities of "distorting" the teaching of [this subject]$_2$.

(52) Wednesday, Energy Secretary Bill Richardson suggested [a new price cap]$_1$ for electricity throughout the Western Sates ... Energy suppliers oppose [a cap]$_2$, saying instead they need incentives to build more generating stations.

(53) [Credit-card]$_1$ issuers have given consumers plenty of reasons in this economic crisis not to pay with [plastic]$_2$.

(54) The Theatre of Palamós will stage next Sunday at 7 PM [the concert entitled "The shawm beyond the cobla"]$_1$. [This concert]$_2$ was created in 2000. Since 2000, [this show]$_3$ has visited different places in Catalonia. The price of seats for [the Sunday concert]$_4$ is 3.01 euros.

(55) But the state defends it as a way to get mothers off drugs, reducing the risk of having [unhealthy babies]$_1$. I went over and looked at [these babies]$_2$ when this case started.

(56) If the United States has officially restored diplomatic relations with [Yugoslavia]$_1$, President Clinton announced the move during his visit to Vietnam, calling the changes in [Yugoslavia]$_2$ remarkable, following the democratic election of President Vojislav Kostunica and the ouster of Slobodan Milosevic.

(57) As a comedian, [Rodney Dangerfield]$_1$ often says [he]$_2$ gets no respect.

(58) [The plant]$_1$ colonized the South of France, from where [it]$_2$ entered Catalonia in the 80s, spreading quickly ... Also, [it]$_3$ presents an important population in the high basin of the Segre River.

(59) For centuries here, [the people]$_1$ have had almost a mystical relationship with Popo, believing the volcano is a god. Tonight, [they]$_2$ fear it will turn vengeful.

(60) The Venezuelan pugilist Antonio Cermeño was stripped of [the super bantamweight interim champion title of the World Boxing Association]$_1$ as he did not meet the requirement of competing for [this crown]$_2$ within the established timeframe ... another Venezuelan, Yober Ortega, will compete for [the vacant crown]$_3$ against the Japanese Kozo Ishii.

APÈNDIX C

Respostes a la tasca de quasi-identitat

Aquest apèndix conté les respostes donades pels sis anotadors en la tasca de quasi-identitat de l'apartat 8.5 utilitzant els fragments inclosos a l'apèndix B. En la tasca es demanava als anotadors que classifiquessin en un dels tipus de relació de (quasi-)identitat (apartat 8.4) els parells de SNs entre claudàtors de cada fragment.

En les tres taules següents, la primera columna mostra entre parèntesis el número del fragment i els números d'identificació dels dos SNs la relació dels quals s'analitza. La resta de les columnes mostra el nombre de vegades que un tipus de quasi-identitat (vegeu la llegenda al principi de cada apartat) va ser assignat a cada parell de SNs (s'indica només si és diferent de 0). Noteu que les files que sumen més de sis són casos en què un o més anotadors va(n) donar múltiples respostes.

C.1 Experiment 1

LLEGENDA

1 No identitat; **2** Identitat; **3** Quasi-identitat;
3A Rol; **3B** Lloc·Agència; **3C** Producte·Productor; **3D** Realització de la informació;
3E Funció numèrica; **3F** Representació; **3Ga** Meronímia–Part·Tot; **3Gb** Meronímia–Conjunt·Membres;
3Gc Meronímia–Porció·Massa; **3Gd** Meronímia–Matèria·Objecte; **3Ha** Interpretació–Selecció;
3Hb Interpretació–Punt de vista; **3Ia** Classe–Més específic; **3Ib** Classe–Més general;
3Ja Funció espaciotemporal–Lloc; **3Jb** Funció espaciotemporal–Temps

Parell	1	2	3A	3B	3C	3D	3E	3F	3Ga	3Gb	3Gc	3Gd	3Ha	3Hb	3Ia	3Ib	3Ja	3Jb
(1)1-2					5					1								
(2)1-2		6								2					1			
(3)1-2						6									1			
(3)1-3		3				3									1			
(3)1-4		3				2									1			
(3)1-5		1				5									1			
(3)2-3		3				3												
(3)2-4		2				4												
(3)2-5						6												
(3)3-4		1				5												
(3)3-5						6												
(3)4-5		2				5												
(4)1-2		5											1					1
(4)1-3		3														1		3
(4)2-3		3											1			1		2
(5)1-2		4											3					
(6)1-2	1	1		4														
(6)1-3		5		1														
(6)2-3	1			5														
(7)1-2	4							2										
(7)1-3		6																
(7)1-4		5												1				
(7)2-3	4							2										
(7)2-4	4							2										
(7)3-4		5												1				
(8)1-2		1							1	5								
(8)1-3		1							1	4								
(8)1-4		1							1	4								
(8)1-5		1							1	5								
(8)2-3		6								1								
(8)2-4		6								1								
(8)2-5		6																
(8)3-4		6																
(8)3-5		6								1								
(8)4-5		6								1								
(9)1-2	5		3															1
(10)1-2		4									1	1	1					
(10)1-3		3									1	1	1		1			
(10)1-4		4									1	1	1					
(10)1-5		3									1		2					
(10)2-3		5														1		
(10)2-4		3								1		1				1		
(10)2-5		2											2		1	1		
(10)3-4		4								1		1						
(10)3-5		3											2		1			
(10)4-5		2										1	2		1			
(11)1-2		2	1					2					4					
(12)1-2		6														2		
(13)1-2		5		1					2									
(14)1-2		4								1								1
(14)1-3		2												1				3
(14)1-4		5																1
(14)1-5		2											1					3
(14)2-3		2																4
(14)2-4		5								1								
(14)2-5													1					5
(14)3-4		1												1				4
(14)3-5		4											1					1
(14)4-5													1					5
(15)1-2		2					1			1					1	2		
(16)1-2									2	4								
(16)1-3	1								1	4								
(16)1-4	1	2								3								
(16)1-5	1	2								3								
(16)2-3	1	2								2					2			
(16)2-4	1								1	3					1			
(16)2-5	1								1	3					1			
(16)3-4		3							1	1				1				
(16)3-5		4							1	1								
(16)4-5		5								1				1				
(17)1-2		4	3															

continua a la pàgina següent

ve de la pàgina anterior

Parell	1	2	3A	3B	3C	3D	3E	3F	3Ga	3Gb	3Gc	3Gd	3Ha	3Hb	3Ia	3Ib	3Ja	3Jb
(17)1-3		1	4										2					
(17)2-3		3	3										2					
(18)1-2		6																
(19)1-2	2						3							2			2	
(20)1-2		5													1			
(20)1-3		2											2		1			2
(20)2-3		2											2		1			2

Taula C.1: Respostes dels anotadors a l'experiment 1

C.2 Experiment 2

LLEGENDA

1 No identitat; **2** Identitat; **3** Quasi-identitat; **3Aa** Metonímia–Rol; **3Ab** Metonímia–Lloc; **3Ac** Metonímia–Organització; **3Ad** Metonímia–Realització informacional; **3Ae** Metonímia–Representació; **3Af** Metonímia–Altres; **3Ba** Meronímia–Part·Tot; **3Bb** Meronímia–Conjunt·Membres; **3Bc** Meronímia–Matèria·Objecte; **3Bd** Meronímia–Superposició; **3Ca** Classe–Més específic; **3Cb** Classe–Més general; **3Da** Funció espaciotemporal–Lloc; **3Db** Funció espaciotemporal–Temps; **3Dc** Funció espaciotemporal–Funció numèrica; **3Dd** Funció espaciotemporal–Rol

Parell	1	2	3Aa	3Ab	3Ac	3Ad	3Ae	3Af	3Ba	3Bb	3Bc	3Bd	3Ca	3Cb	3Da	3Db	3Dc	3Dd
(21)1-2		2		5														
(22)1-2		2											4					
(23)1-2		6																
(23)1-3		6																
(23)1-4	2								3	2								
(23)2-3		6																
(23)2-4	2								3	2								
(23)3-4	2								3	2								
(24)1-2						6												
(24)1-3		6																
(24)1-4		6																
(24)2-3						6												
(24)2-4						6												
(24)3-4		6																
(25)1-2	1	1								4			1					
(25)1-3										3		2	2					
(25)2-3												5	1					
(26)1-2		1											1			2	3	
(26)1-3																2	4	
(26)2-3																2	4	
(27)1-2		1	4				2											
(27)1-3		1					5											
(27)1-4		5					1											
(27)1-5		5					1											
(27)2-3		3					3											
(27)2-4			4				3											
(27)2-5			4				3											
(27)3-4							6											
(27)3-5							6											
(27)4-5		6																
(28)1-2	3							2						1				
(29)1-2				4					2									
(30)1-2	1	5			2													
(30)1-3	1	5			2													
(30)1-4	1	2			2					2								
(30)2-3		6																
(30)2-4		2			1					4								
(30)3-4		2			1					4								
(31)1-2					4										1	1		
(32)1-2		1														5		
(33)1-2		2										2	3					
(34)1-2		6							1					1		1		
(35)1-2				6														
(36)1-2		2										1	1	2				
(37)1-2		2										4						
(38)1-2	3	1	2															
(39)1-2				1					5									
(39)1-3				3					3									
(39)1-4		6																
(39)2-3	1			1					3	1								
(39)2-4				1					5									
(39)3-4				3					4									
(40)1-2		1							1	5								

Taula C.2: Respostes dels anotadors a l'experiment 2

C.3 Experiment 3

LLEGENDA

1 No identitat; **2** Identitat; **3** Quasi-identitat;

3Aa Metonímia–Rol; **3Ab** Metonímia–Lloc; **3Ac** Metonímia–Organització;

3Ad Metonímia–Realització informacional; **3Ae** Metonímia–Representació; **3Af** Metonímia–Altres;

3Ba Meronímia–Part·Tot; **3Bb** Meronímia–Matèria·Objecte; **3Bc** Meronímia–Superposició;

3Ca Classe–Més específic; **3Cb** Classe–Més general; **3Da** Funció espaciotemporal–Lloc;

3Db Funció espaciotemporal–Temps; **3Dc** Funció espaciotemporal–Funció numèrica;

3Dd Funció espaciotemporal–Rol

Parell	1	2	3Aa	3Ab	3Ac	3Ad	3Ae	3Af	3Ba	3Bb	3Bc	3Ca	3Cb	3Da	3Db	3Dc	3Dd
(41)1-2	1														6		
(42)1-2				6													
(43)1-2	6																
(43)1-3	6																
(43)1-4	6																
(43)2-3		1									6						
(43)2-4		3									3						
(43)3-4		1									6						
(44)1-2		2						4									
(45)1-2		2									3	1	1				
(45)1-3		1									5						
(45)1-4		1									5						
(45)2-3		1									5						
(45)2-4		1									5						
(45)3-4		2									4						
(46)1-2		1			5												
(47)1-2		5										1					
(47)1-3		6															
(47)2-3		5											1				
(48)1-2		3			3										1		
(48)1-3		3			3										2		
(48)2-3		3			3										2		
(49)1-2			2				4					1					
(50)1-2				5					1								
(51)1-2		2			1	3											
(52)1-2		3											4				
(53)1-2										5			1				
(54)1-2		5													1		
(54)1-3						4								2			
(54)1-4		2				4									2		
(54)2-3		2				4											
(54)2-4						1						1		1	5		
(54)3-4														3	5		
(55)1-2	1	1									1	4					
(56)1-2		1		5													
(57)1-2		6															
(58)1-2		4												5			
(58)1-3		3												5			
(58)2-3		3												5			
(59)1-2	1										4				4		
(60)1-2		3											3				
(60)1-3		1													4		1
(60)2-3		2													4		1

Taula C.3: Respostes dels anotadors a l'experiment 3