Coreference:
Theory, Annotation, Resolution and Evaluation

by

Marta Recasens Potau

A Dissertation Presented to the
Doctoral Program in *Linguistics and Communication*,
Department of Linguistics,
University of Barcelona,
in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

under the supervision of

Dr. M. Antònia Martí Antonín
University of Barcelona

Dr. Eduard Hovy
ISI – University of Southern California

University of Barcelona
September 2010

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

Counting words isn't very revealing if you aren't listening to them, too.

– Geoffrey Nunberg

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

Usage factors reveal language as a natural, organic social instrument,
not an abstract logical one.

– Joan Bybee

Language, Usage and Cognition (2010:193)

To my loved ones, Mercè, Eduard, Elm and Mark, for being there.

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.

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.

Acknowledgments

To cook my thesis, I have had the opportunity to work under the supervision of two very talented and enthusiastic chefs, M. Antònia Martí and Ed Hovy. I am deeply grateful to them for revealing to me the secret ingredients that make research fascinating, and for always being present—physically or electronically—to teach me treasured recipes that have become part of my cooking style. Over the past four years, they have guided me through all the steps required to prepare an appetizing thesis: planning, creativity, decision making, elegance, patience, artful serving, and careful and slow cooking. Thank you for your dedication to help me master this art and for being much more than advisors.

Many thanks also to the members of my committee, Costanza Navarretta, Massimo Poesio, and Mariona Taulé, for accepting to be on my advisory committee. My sincere gratitude for your feedback and the ideas you contributed at various moments during the cooking process that have helped make this a nutritious experience.

I have been very lucky to meet and interact with other extraordinarily skillful chefs who have suggested excellent recipes to try or have given sound advice on how to improve a dish. I am indebted to Jerry Hobbs, Horacio Rodríguez, Mihai Surdeanu, Lluís Màrquez, Ruslan Mitkov, Vincent Ng, Véronique Hoste, Antal van den Bosch, Olga Uryupina, and Manu Bertran. A very huge thank you goes to Edgar Gonzàlez, who has always been willing to help me debug code and when I got stuck on some computer problem. Thank you for teaching me how to taste Java.

I also feel very fortunate of having been surrounded by fantastic colleagues at two of the finest kitchens, the Linguistics Department at the University of Barcelona

and the Information Sciences Institute at the University of Southern California. Warm thanks go to my officemates Marta Vila and Aina Peris with whom I have shared countless hours of friendship, hard work, fun, and tears. I have greatly enjoyed the company of and would like to thank them all for creating an inspiring working atmosphere: 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, and Rita Zaragoza. I also extend my sincere gratitude to my "conference friends" Constantin Orasan, Laura Hasler, and Marta Ruiz Costa-Jussà, with whom I had fruitful discussions while enjoying Bulgarian, Czech, Moroccan, and Indian food. I want to single out Constantin for taking time to read my dissertation and making a number of valuable suggestions.

I would like to thank my dearest parents, from whom I learned to savor life's little moments, for being my very first teachers and for always encouraging me to pursue my intellectual interests. Thanks also to my brother Elm, who knows real cooking, for making me the most delicious black rice that kept my energy levels up. I want to thank my great friends for not losing the habit of getting together for dinner, and making an effort to understand what I spent the hours working on. A special thanks goes to Laura, Joana, Sara, Lali, Cristina, Marc, Carlota, Marta, Maria, Laia, Manel, Antonio, Tina, and Martí (for all those glasses of *orxata*!). Last but certainly not least, thank you Mark for adding that magic touch of joy and love and for bringing your flavor into my life.

⋆ ⋆ ⋆

This work was supported by an FPU Grant (AP2006-00994) from the Spanish Ministry of Education and Science.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

This thesis is about coreference. It is the story of a project that set out to annotate a corpus with coreference relations to train the first machine learning coreference resolution systems of Spanish and Catalan, but ended up developing an alternative theoretical view of coreference. This view grew from the felt need to revisit the definition of coreference, to reconsider what can be solved automatically and to rethink evaluation criteria. The whys and wherefores of this turn are covered in the next 200 pages.

The structure of this thesis is intentionally chronological to capture the four-year development that shaped the arguments I put forward. The original idea gradually changed, leading to the completion of this thesis. Thus, I begin by placing the readers in the same conditions as when I started, to follow the logical progression from beginning to end.

1.1 Point of departure

The point of departure for this thesis was the problem of coreference resolution, one of the challenging tasks for Natural Language Processing (NLP). It is usually defined as either "the problem of identifying which noun phrases (NPs) or mentions refer to the same real-world entity in a text or dialogue" (Ng, 2009; Stoyanov et al., 2009; Finkel and Manning, 2008) or, in slightly different terms, "the task of grouping all the mentions of entities in a document into equivalence classes so that all the mentions in a given class refer to the same discourse entity" (Bengtson and Roth, 2008; Denis and Baldridge, 2009). Accordingly, mentions 1, 2 and 3 in (1) are said to corefer, as all three refer to Eyjafjallajökull.[1]

[1] Because coreference is a discourse phenomenon, it will usually not be possible to spare much text length in the examples provided throughout.

(1) [The Eyjafjallajökull volcano, one of Iceland's largest,]$_1$ had been dormant for nearly two centuries before returning gently to life in the late evening of March 20, 2010, noticeable at first by the emergence of a red cloud glowing above the vast glacier that covers [it]$_2$. In the following days, fire fountains jetted from a dozen vents on [the volcano]$_3$, reaching as high as 100 meters.[2]

Note that different linguistic expressions are used—a proper noun, a pronoun, and a definite NP. This, however, is not a requirement for coreference: mentions 1, 2 and 3 in (2) would also be coreferent with the only difference being a loss of discourse cohesion.

(2) [Eyjafjallajökull]$_1$ had been dormant for nearly two centuries before returning gently to life in the late evening of March 20, 2010, noticeable at first by the emergence of a red cloud glowing above the vast glacier that covers [Eyjafjallajökull]$_2$. In the following days, fire fountains jetted from a dozen vents on [Eyjafjallajökull]$_3$, reaching as high as 100 meters.

Coreference, like identity, is defined as an either-or relation: two mentions are either coreferent (i.e., they have identical referent) or non-coreferent (i.e., they have different referent). Early work on coreference resolution derived from the sister task of anaphora resolution (Mitkov, 2002), which involves solving the reference of (anaphoric) pronouns and definite NPs whose interpretation depends on a previous expression, i.e., identifying their antecedent in the text. Although related, coreference resolution goes one step further as it requires resolving the reference of all mentions in the text (pronouns, proper nouns, definite and indefinite NPs, etc.), including those that do not depend on another expression for their interpretation.

We, as language users, can quickly and unconsciously work out the reference of every linguistic expression, linking the information provided by those that refer to the same entity. However, the underlying process of how this is done is yet unclear. The question of making explicit in a systematic way the knowledge behind such practices remains a difficult one, and thus the challenge that coreference resolution poses to NLP. There is nonetheless a strong interest in automatically identifying coreference links as they are key to "understand" a text and so they are needed by NLP applications such as information extraction (McCarthy and Lehnert, 1995), text summarization (Azzam et al., 1999; Steinberger et al., 2007), question answering (Morton, 2000; Vicedo and Ferrández, 2006) and machine translation, where the antecedent of a pronoun has to be identified before it can be translated. Coreference links are also useful for other tasks like sentiment analysis (Nicolov et al., 2008), textual entailment (Mirkin et al., 2010; Abad et al., 2010), citation matching and databases (Wick et al., 2009), machine reading (Poon et al., 2010), for learning narrative schemas (Chambers and Jurafsky, 2008) and for recovering implicit

[2] The New York Times (April 20, 2010).

arguments (Gerber and Chai, 2010; Ruppenhofer et al., 2010).

Adding to the existing research on coreference resolution, this thesis arose from the interest in improving state-of-the-art coreference resolution by making extensive use of machine learning in a linguistically informed way. My aim was to uncover underlying patterns of coreference relations and to generalize how different linguistic cues interact and are weighed against each other. In addition, being English the primary focus of NLP research, I was also motivated by the need to develop language resources for Spanish and Catalan such as a coreferentially annotated corpus and a coreference resolution system.

Corpus annotation was seen as an opportunity to bring new perspectives in coreference resolution. My working hypothesis was that annotating coreference relations by placing emphasis on both the quantity of the annotated data and the quality of this annotation would have an immediate positive effect on the model learned by machine learning methods and, in turn, on the performance of coreference resolution systems. Apart from the quantitative aspect, which mainly implied a time-consuming and expensive procedure, the challenge of annotation came down to gaining a full understanding of the coreference phenomenon:

> Statistical models of anaphora resolution so far have only scratched the surface of the phenomenon, and the contributions to the linguistic understanding of the phenomenon have been few. (Poesio et al., forthcoming:85)

> The notions of coreference and anaphora are difficult to define precisely and to operationalize consistently. Furthermore, the connections between them are extremely complex. (Stoyanov et al., 2009:657)

In my approach to the problem, I conducted a study of empirical cases and considered language-specific properties in order to define a linguistically accurate coding scheme (Chapter 2). I wanted annotation to escape from computational demands such as those imposed by the MUC guidelines (Hirschman and Chinchor, 1997), which make the coreference task definition dependent on supporting, as the first priority, the MUC information extraction tasks. At a later stage, and at a more theoretical level, I was driven by the challenges of comparing coreference and paraphrase, another phenomenon with which it bears some resemblance (Chapter 7), and of properly defining "identity of reference," a notion taken for granted but conceptually very complex (Chapter 8): Do *Postville* and *the old Postville* refer to the same entity? Is *the broken glass* the same as *the unbroken piece*? These have been longly-debated questions captured by so-called identity paradoxes such as Heraclitus' river and Theseus's ship.

Armed with a gold-standard corpus—annotated not only with coreference relations but also with morphological, syntactic and semantic information—I faced the challenge of resolution. On the one hand, I intended to draw on the annotation experience to define linguistically motivated features (Chapter 3). My interest was in exploring the feature space with the help of machine learning techniques, which are known for their efficiency in handling large numbers of features. On the other hand, the limitations of mention-pair models was to be addressed by designing an entity-based system that would take whole entity clusters into consideration (Chapter 4). In this way, more linguistic information could be used to decide whether or not to add a mention into an entity.

As soon as the first performance scores of the prototype system presented in this thesis were obtained, I met the challenge of evaluation. Although several metrics exist to measure the performance of coreference resolution systems (Vilain et al., 1995; Bagga and Baldwin, 1998; Luo, 2005), there is still no agreement on a standard. Criteria needed to be established to assess the quality of a coreference output: What should be more rewarded in (1), linking as coreferent mentions 1, 2 and 3 plus *the vast glacier*, or linking only mentions 1 and 3? Or what should be more penalized, linking mention 1 and *the emergence of a red cloud glowing above the vast glacier that covers it*, or linking *the late evening of March 20, 2010* and *the vast glacier*?

Despite the many published results in the literature, different assumptions in the evaluation methodology hinder an appropriate comparison of state-of-the-art performance scores. I applied several criteria to analyze the pros and cons of the currently used coreference metrics (Chapter 5). The official SemEval shared task (Chapter 6) provided further insight into the challenges posed by evaluation and system comparisons (Chapter 4).

1.2 Thesis outline

The present thesis consists of a collection of seven papers sandwiched between an introductory and a concluding chapter that provide the necessary glue to make the thesis constitute a whole. The seven papers are the following:

Part I: Corpus Annotation with Coreference

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

Part II: Coreference Resolution and Evaluation

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

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

4. Recasens, Marta and Eduard Hovy. To appear. 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, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the ACL 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 1–8, Uppsala, Sweden.

Part III: Coreference Theory

6. Recasens, Marta and Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.

7. Recasens, Marta, Eduard Hovy, and M. Antònia Martí. In revision. Identity, non-identity, and near-identity: Addressing the complexity of coreference. Submitted to *Lingua*.

The first six papers have been or soon will be published in peer reviewed journals or conference proceedings, and the last one is currently under review. They are co-authored by either one or both of my advisors, with two exceptions: paper 5 required collaboration between several research groups to organize the shared task, and paper 6 resulted from joint work with another graduate student at the University of Barcelona who researches paraphrase. In all cases I am listed as the first author. The papers are reprinted here reformatted to make the typography of the thesis consistent, and references and appendices are integrated in a single bibliography and appendix section at the end.

The thesis is organized in three parts. The first part, comprising the next chapter, focuses on the annotation of corpora with coreference information using the case of the Spanish and Catalan AnCora corpora. The second part of the thesis, comprising Chapters 3 through 6, discusses my experience in developing and evaluating coreference resolution systems. More specifically, I concentrate on the feature set definition, the interdependence between system and corpus parameters, the behavior of coreference evaluation metrics and the SemEval shared task that set up a testbed to compare different systems. The third part of the thesis, corresponding to Chapters 7 and 8, provides a better theoretical understanding of coreference by first delimiting the scope of the concept as opposed to that of paraphrase, and by secondly presenting a continuum approach to coreference that introduces the notion of *near-identity* and that opens up a new avenue for future research.

In the remainder of this chapter, I first define key terms in coreference research, and review previous work to put my work into perspective. Then, I provide the links and connections between the seven papers by making explicit how the outcomes of the different stages influenced each other and led me to pursue new directions. Finally, I recapitulate the major contributions to the state of the art.

1.3 Key terminology

To begin with, and in order to clarify notions commonly used in the field of coreference resolution, I give a brief explanation and note some remarks of the terms relevant to this study.

Mention and **entity** The MUC and ACE programs[3] (Hirschman and Chinchor, 1997; Doddington et al., 2004) have popularized these two terms in the field of coreference resolution. As defined by ACE, an *entity* is "an object or set of objects in the world." In addition, the ACE program restricts entities to a few specific types (person, organization, location, etc.). A *mention*, on the other hand, is a textual reference to an entity. In other words then, an entity corresponds to the collection of mentions referring to the same object. A couple of observations are in place. First, mentions are referential NPs, which means that they exclude expletive pronouns (e.g., *It is raining*), attributive or predicative NPs (e.g., *He is a member of the company*), and idiomatic NPs (e.g., *It's raining cats and dogs*). Some approaches, however, assume a broader interpretation and use "mention" as a synonym for "NP." Second, the claim that entities are *in the world* requires further consideration, as discussed next.

Discourse model and **discourse entity/referent** Discourse representation theories, concerned with the representation of discourse and the processes involved in the comprehension and production of discourse, consider that linguistic reference is not a mapping from linguistic expressions to the real world but to constructs in the discourse model built as the text progresses. In the following excerpt, Prince (1981:235) summarizes the basic notions:

> Let us say that a text is a set of instructions from a speaker to a hearer on how to construct a particular *discourse model.* The model will contain *discourse entities*, attributes, and links between entities. A discourse entity is a discourse-model object, akin to Karttunen's (1976) *discourse referent*; it may represent an individual (existent in the real world or not), a class of individuals, an exemplar, a substance, a concept, etc. Following Webber (1979) entities may be thought of as hooks on which to hang attributes. All discourse entities in a discourse model are represented by NPs in a text, though not all NPs in a text represent discourse entities.

Henceforth, by "entity" I mean "discourse entity."

[3] The Message Understanding Conferences (MUC) and the Automatic Content Extraction (ACE) evaluation were initiated and financed by the DARPA agency of the U.S. Department of Defense, and the National Institute of Standards and Technology of the U.S. Department of Commerce, respectively, to encourage the development of new and better methods of information extraction.

Anaphor and **antecedent** An *anaphor* designates a linguistic expression that depends on a previous one (its *antecedent*) for its interpretation. The pronoun *they* has *Airlines* as an antecedent in *Airlines are still uncertain about when they can return to a regular schedule*. While anaphora is a textual relation that requires the reader to go back in the text to interpret an empty (or almost empty) textual element, coreference occurs at the referential level. The terms "anaphor" and "antecedent" properly belong to the domain of anaphora resolution, but they are also used in coreference resolution in two ways: a correct and an incorrect one. They are correctly used to refer to antecedent-anaphor pairs that are part of a coreference chain, like *airlines* and *they*, or *Eyjafjallajökull* and *the volcano* in (1) above; whereas they are incorrectly used to refer indistinctly to any mention pair of a coreference chain, like *the Eyjafjallajökull volcano* and *Eyjafjallajökull*, as the latter does not require the former to be interpreted.

Discourse-new/first mention and **subsequent mention** The counterparts of "antecedent" and "anaphor" in the field of coreference resolution are *discourse-new* or *first* mention and *subsequent* mention, respectively. A discourse-new or first mention introduces a (new) entity in the text; it is thus its first mention, like *Eyjafjallajökull*$_1$ in (2) above. A subsequent mention, in contrast, is any later mention of an entity already introduced in the text, like *Eyjafjallajökull*$_2$ and *Eyjafjallajökull*$_3$ in (2). Notice that subsequent mentions can be, but are not always, anaphoric. Nevertheless, the term "non-anaphoric" is often incorrectly used to mean 'discourse-new,' and "anaphoric" incorrectly used to mean 'subsequent mention.'

Singleton and **multi-mention entity** Depending on the size of the entity, i.e., the number of mentions it contains, it is convenient to distinguish between *singletons* (or *singleton entities*) if they have only one mention, and *multi-mention entities* if they have two or more mentions. The former are also called *isolated* mentions, as they make an isolated reference to an entity. Coreference relations, then, are only possible within multi-mention entities.

Pairwise/mention-pair model and **entity-based/entity-mention model** There are two basic kinds of coreference resolution models. On the one hand, *pairwise* (or *mention-pair*) models work in terms of pairs of mentions, classifying two mentions as either coreferent or non-coreferent, and then combining all the pairwise decisions to partition the document mentions into coreference chains. On the other hand, *entity-based* (or *entity-mention*) models are meant to improve the classification by determining not the probability that a mention corefers with a previous mention but the probability that a mention refers to a previous entity, i.e., a set of mentions already classified as coreferent. Thus, the latter often employ clustering strategies.

Link-based measure and **class-based measure** In parallel to the two types of resolution strategies, coreference systems can be evaluated using two different types of measures: *link-based* performance metrics are based on the number of correct, missing and spurious links (i.e., pairs of coreferent mentions) identified by the system, whereas *class-based* performance metrics treat entities as clusters and take into consideration not only multi-mention entities but also singletons.

End-to-end system and **coreference module** The term *coreference resolution system* is ambiguous between an *end-to-end system*, a system capable of determining coreference on plain text, that is, of identifying the mentions and their boundaries automatically before predicting coreference relations, and a *coreference module*, which strictly identifies coreference relations assuming that the data contain gold-standard linguistic information at various levels (mention boundaries, PoS, parse trees, etc.).

True/gold mentions and **system mentions** The set of mentions contained in the gold standard, produced by human expert, are referred to as *true* or *gold mentions*, as opposed to the set of mentions contained in the system output, which are called *system mentions*. As a follow-up to the previous paragraph, true and system mentions do not usually coincide when using an end-to-end coreference system, but they do in the case of a coreference module.

1.4 Related work

A number of good surveys (Poesio et al., forthcoming; Ng, 2010; Ng, 2003) provide a wide overview of computational approaches to coreference and anaphora resolution. This section is meant to give the reader not another extensive account, but a compact overview of the previous and latest research on the main issues related to this thesis in order to supplement and bind together the “Background” sections already provided in each paper.

From a thematic perspective, the subject matter can be broken down into the four areas that roughly correspond to the key steps followed in developing a coreference resolution system: (i) corpus creation, (ii) learning features, (iii) classification and clustering, and (iv) evaluation. I will deal with each of them in turn, highlighting the main trends and milestones. The reader is referred to the original papers for details.

1.4.1 Corpus creation

Research on coreference resolution with a view to practical applications requires corpora annotated with coreference information for two main reasons: (i) to train machine-learning systems, and (ii) to test automatic systems on large-scale data.

In addition, corpora annotated with coreference information are valuable in usage-based linguistics to study language based on authentic usage data. Annotating coreference, however, is not a trivial task, and there have been numerous proposals for coding schemes. In fact, each corpus usually defines its own scheme as there is no agreed-upon standard. From a global viewpoint, it is possible to distinguish between application-oriented and linguistically-oriented approaches.

Application-oriented approaches The MUC and ACE corpora (Hirschman and Chinchor, 1997; Doddington et al., 2004) were specifically designed for shared tasks on information extraction. Consequently, annotation decisions—like the set of coreferent elements or the scope of the identity relation—are subordinated to the needs of the tasks even if it is at the cost of linguistic accuracy. Thus, they do not annotate the entire NP with all its modifiers but only up to the head, and verbal or clausal mentions are ignored. Moreover, ACE limits the set of NPs to seven semantic types (relevant to the information-extraction domain): person, organization, geo-political entity, location, facility, vehicle, and weapon. In a similar vein, MUC and ACE tailor the definition of "identity of reference" to suit the needs of information extraction, treating nominal predicates and appositions also as coreferent. For this reason, MUC has been trenchantly criticized by van Deemter and Kibble (2000) for conflating "elements of genuine coreference with elements of anaphora and predication in unclear and sometimes contradictory ways."

Linguistically-oriented approaches As a response to the MUC approach, the MATE meta-scheme, those derived from it (GNOME, ARRAU [Poesio, 2004a; Poesio and Artstein, 2008]) as well as others like OntoNotes (Pradhan et al., 2007b) aim to create corpora not for a specific task but for research on coreference at large. They allow a wider range of syntactic types (i.e., mentions other than NPs) and nominal mentions map onto NPs including all their modifiers. MATE goes beyond and contemplates linguistic phenomena typical of Romance languages such as elliptical subjects and incorporated clitics. An explicit separation is kept between the identity relation on one hand and the appositive relation (for nominal predicates and appositive phrases) on the other. In addition, MATE suggests the annotation of relations other than identity (set membership, subset, possession, bound anaphora, etc.) as well as ambiguities.

Despite the distinction between these two directions, the diversity of existing schemes and idiosyncratic tags incorporated by different corpora is a reflection of the lack of a general and satisfactory theory of coreference that does not solely rely on the simple "identity of reference" definition. The choice of corpus when developing or testing a coreference system is not a minor issue, and the way corpora are annotated has greatly determined the design and architecture of systems.

Language resources Table 1.1 summarizes the over-25k-word corpora that have been annotated with coreference information, with newspaper texts as the domi-

Corpus	**Reference**	**Language**	**Genre**	**Size**
ACE-2	Doddington et al. (2004)	English	News	180k
ACE-2003, ACE-2004, ACE-2005		Arabic, Chinese, English	News, weblogs	100-350k
ACE-2007		Arabic	News, weblogs	220k
		Chinese	News, weblogs	250k
		English	News, dialogues, weblogs, forums	300k
		Spanish	News	200k
AnATAr	Hammami et al. (2009)	Arabic	News, textbook, novel, technical manual	77k
AnCora-Ca	Recasens and Martí (2010)	Catalan	News	400k
AnCora-Es		Spanish	News	400k
ARRAU	Poesio and Artstein (2008)	English	Dialogues, spoken narratives, news, GNOME	100k
C-3	Nicolae et al. (2010)	English	News, aptitude tests	75k
COREA	Hendrickx et al. (2008)	Dutch	News, spoken language, encyclopedia entries	325k
DAD	Navarretta (2009b)	Danish, Italian	News, law texts, narratives	25k
GNOME	Poesio (2004a)	English	Museum labels, leaflets, dialogues	50k
I-CAB	Magnini et al. (2006)	Italian	News	250k
KNACK-2002	Hoste and De Pauw (2006)	Dutch	News	125k
LiveMemories	Rodríguez et al. (2010)	Italian	News, Wikipedia, dialogues, blogs	150k
MUC-6	Grishman and Sundheim (1996)	English	News	30k
MUC-7	Hirschman and Chinchor (1997)	English	News	25k
NAIST Text	Iida et al. (2007)	Japanese	News	970k
NP4E	Hasler et al. (2006)	English	News	50k
Switchboard	Calhoun et al. (2010)	English	Telephone conversations	200k
OntoNotes 2.0	Pradhan et al. (2007a)	English	News	500k
		Arabic	News	100k
		Chinese	News	550k
Potsdam Commentary	Stede (2004)	German	News	33k
PDT 2.0	Kučová and Hajičová (2004)	Czech	News	800k
TüBa-D/Z	Hinrichs et al. (2005)	German	News	800k
Venex	Poesio et al. (2004a)	Italian	News, dialogues	40k

Table 1.1: Summary of the largest coreferentially annotated corpora

nant genre. For the sake of completeness, I also include the two corpora that are a contribution of this thesis (AnCora-Ca and AnCora-Es).[4] They clearly fill the gap of resources for both Catalan and Spanish. No Catalan data annotated with coreference existed before AnCora-Ca, and AnCora-Es overcomes the Spanish ACE-2007 corpus not only in size but also in the limitations imposed by ACE-type entities. Given that the focus of the present thesis is coreference, Table 1.1 does not include corpora that are only annotated with anaphoric pronouns like the Spanish Cast3LB corpus (Navarro, 2007). The ongoing ANAWIKI annotation project aims to collect large amounts of coreference data for English via a Web collaboration game called *Phrase Detectives* (Poesio et al., 2008). Although the most number of resources available are for English, the past years have seen a growing interest in providing other languages with coreferentially annotated data.

1.4.2 Learning features

Before large data sets annotated with coreference information became available in the mid-1990s, the immediate ancestors to today's machine-learning coreference systems were pronominal anaphora resolution systems that relied on a set of hand-crafted rules (Hobbs, 1978; Rich and LuperFoy, 1988; Carbonell and Brown, 1988; Alshawi, 1990; Kameyama, 1998; Tetreault, 2001; Palomar et al., 2001), especially in the form of constraints and preferences.

Constraints and preferences Given a pronoun to resolve, constraints rule out incompatible antecedents, whereas preferences score the remaining candidates in order to select the best antecedent. They are based on information from different linguistic levels, as displayed in the first row of Table 1.2, although the biggest emphasis is on syntax (Hobbs, 1978; Carbonell and Brown, 1988) and Centering theory (Kameyama, 1998; Tetreault, 2001). There was, however, an increasing tendency to replace knowledge-rich systems with knowledge-poor ones that would do without semantic and world knowledge (Lappin and Leass, 1994) or, even more, without assuming full syntactic parsing (Kennedy and Boguraev, 1996; Baldwin, 1997; Mitkov, 1998).

Heuristics of this type served to capture the most important rules governing antecedent–pronoun relations. However, the greater level of complexity of coreference resolution accounts in part for the shift from heuristic to machine-learning approaches in the past decade. Applying machine learning to large-scale data sets enables the ordering and weighing of large feature sets more efficiently than rule-based heuristic approaches. Both Aone and Bennett (1995) and McCarthy and Lehnert (1995) report that their rule-based classifiers are outperformed by their learning-based counterparts.

[4] The name *AnCora* (ANnotated CORporA) is used generically to designate the Spanish and Catalan corpora with all their layers of annotation. Suffixes can be attached to refer to a specific part: *AnCora-Ca* designates the Catalan portion, *AnCora-Es* designates the Spanish portion, *AnCora-CO* designates the coreference annotations of the corpora, etc.

Constraints and preferences for pronoun resolution (Hobbs, 1978; Rich and LuperFoy, 1988; Carbonell and Brown, 1988; Alshawi, 1990; Lappin and Leass, 1994; Kennedy and Boguraev, 1996; Baldwin, 1997; Mitkov, 1998; Kameyama, 1998; Tetreault, 2001)	**1.** Gender agreement, **2.** Number agreement, **3.** Binding constraints, **4.** m_i is subject, **5.** Animacy, **6.** Selectional constraints, **7.** Mention embedded within a quantified or negated structure, **8.** Case-role parallelism, **9.** Syntactic parallelism, **10.** m_i is in a topicalized structure, **11.** Sentence distance, **12.** Recency, **13.** Grammatical role, **14.** Person agreement, **15.** Frequency of mention, **16.** The postconditions of the action containing m_i violate the preconditions of the action containing m_j, **17.** Mention is embedded, **18.** Mention is in an existential construction, **19.** Centering constraints, **20.** m_i is definite, **21.** m_i is the first NP in the sentence, **22.** m_i is the object of verbs such as *discuss, present, illustrate, describe*, etc., **23.** m_i is in the heading of the section, **24.** Mention is not part of a prepositional phrase, **25.** Mention is a domain term [...]
Basic coreference feature set (Soon et al., 2001)	**1.** m_i is a pronoun, **2.** m_j is a pronoun, **3.** m_j is a definite, **4.** m_j is a demonstrative, **5.** m_i and m_j are proper names, **6.** String match (without determiners), **7.** Number agreement, **8.** Gender agreement, **9.** Semantic class agreement, **10.** m_j is an appositive of m_i, **11.** One mention is an alias of the other, **12.** Sentence distance
Extended coreference feature set (Ng and Cardie, 2002b)	**1.** m_i and m_j are pronominal/proper names/non-pronominal and the same string, **2.** The words of m_i and m_j intersect, **3.** The prenominal modifiers of one mention are a subset of those of the other, **4.** m_i and m_j are proper names/non-pronominal and one is a substring of the other, **5.** m_i and m_j are definites, **6.** m_i and m_j are embedded, **7.** m_i and m_j are part of a quoted string, **8.** m_i is a subject, **9.** m_j is a subject, **10.** m_i and m_j are subjects, **11.** m_i and m_j match in animacy, **12.** m_i and m_j have the same maximal NP projection, **13.** m_j is a nominal predicate of m_i, **14.** m_i is an indefinite and not appositive, **15.** m_i and m_j are not proper names but contain mismatching proper names, **16.** m_i and m_j have ancestor-descendent relationship in WordNet, **17.** WordNet distance, **18.** Paragraph distance [...]
Additional features (Strube et al., 2002; Luo et al., 2004; Nicolae and Nicolae, 2006; Ponzetto and Strube, 2006; Uryupina, 2006; Ng, 2007; Yang and Su, 2007; Bengtson and Roth, 2008)	**1.** Minimum edit distance between m_i and m_j strings, **2.** Head match, **3.** Word distance, **4.** Mention distance, **5.** One mention is an acronym of the other, **6.** Pair of actual mention strings, **7.** Number of different capitalized words in two mentions, **8.** Semantic role, **9.** WordNet similarity score for all synset pairs of m_i and m_j, **10.** The first paragraph of the Wikipedia page titled m_i contains m_j (or vice versa), **11.** The Wikipedia page titled m_i contains an hyperlink to the Wikipedia page titled m_j (or vice versa), **12.** Saliency, **13.** One mention is a synonym/antonym/hypernym of the other in WordNet, **14.** m_i and m_j appear within two words of a verb of diction, **15.** Modifiers match, **16.** Aligned modifiers relation, **17.** Semantic similarity, **18.** Parse tree path from m_j to m_i [...]
Cluster-level features (Luo et al., 2004; Daumé III and Marcu, 2005; Ng, 2005; Culotta et al., 2007; Poon and Domingos, 2008; Yang et al., 2008; Rahman and Ng, 2009)	**1.** Feature X is true of any pair, **2.** All pairs share a feature X, **3.** The majority of pairs share a feature X, **4.** Feature X is false of any pair, **5.** All mention pairs are predicted to be coreferent, **6.** Most mention pairs are predicted to be coreferent, **7.** Decayed density, **8.** Entity to mention ratio, **9.** Size of the hypothesized entity chain, **10.** Count of how many NPs are of each mention type, **11.** Probability that a pair has incompatible gender values [...]

Table 1.2: Summary of seminal coreference feature sets (m_i and m_j stand for two different mentions where $i < j$)

Feature vectors In the classic supervised learning setup (see 1.4.3 below), learning instances are created by pairing two mentions m_i and m_j, and labeling them as either true/coreferent (*positive instance*) or false/non-coreferent (*negative instance*): $\langle m_i, m_j, boolean\rangle$ is true if and only if m_i and m_j are coreferent. Pairs $\langle m_i, m_j\rangle$ are represented by a feature vector consisting of unary features (i.e., information about one of the mentions, e.g., its number) and binary features (i.e., information about the relation between the two mentions, e.g., number agreement). Table 1.2 summarizes the learning features most frequently used (mainly for English), many of which borrow from constraints and preferences. Up to this date, most coreference resolution systems (Bengtson and Roth, 2008; Denis and Baldridge, 2009; Stoyanov et al., 2010) have been modeled after Soon et al.'s (2001) limited but successful feature set improved with the extension by Ng and Cardie (2002b). One of the features that has emerged as the most successful is the appositive one (Soon et al., 2001; Poon and Domingos, 2008).

Additional features While the morphosyntactic and surface features reported in Table 1.2 succeed in solving a majority of coreference relations, they seem to reach a limit, especially in the case of definite NPs and proper names (Vieira and Poesio, 2000; Haghighi and Klein, 2009), that can only be surpassed with the use of deep semantic and world knowledge. The latest models have tried to provide a useful approximation to such knowledge by drawing semantic patterns from resources like WordNet and the World Wide Web (Ponzetto and Strube, 2006; Uryupina, 2006; Ng, 2007), but any improvement, although significant, is small. In this regard, Kehler et al. (2004) point out that predicate-argument statistics mined from naturally-occurring data do not improve performance (for pronoun resolution), as the cases for which statistics hurt are potentially more damning than those for which they help.

Cluster-level features A promising way of incorporating more knowledge without getting entangled in the construction of expensive resources seems to be the design of more global models with room for cluster-level features that make it possible to take into consideration not only two but all the mentions of a (partial) entity (Luo et al., 2004; Culotta et al., 2007; Poon and Domingos, 2008). The design of such features, however, is a complex matter, and most of them derive directly from the old pairwise ones. Apart from incorporating new knowledge sources to strengthen the feature set, performance can also increase substantially with feature selection (Ng and Cardie, 2002b; Hoste, 2005) and training instance selection (Harabagiu et al., 2001; Ng and Cardie, 2002b; Uryupina, 2004; Hoste, 2005), although less attention is usually paid to these. The focus of research has largely shifted from incorporating new features to applying new resolution models, as discussed in 1.4.3.

Languages other than English Only recently has the validity of the coreference features been tested for languages other than English at the same time that coreferentially annotated corpora have become available for these languages. See, for instance, Hoste (2005) for Dutch, Versley (2007) or Klenner and Ailloud (2009) for German, Poesio et al. (2010) for Italian, and Nilsson (2010) for Swedish. Prior to the research reported in this thesis, the case of Spanish and Catalan remained virtually unexplored—except for a few rule-based pronoun resolution systems for Spanish (Palomar et al., 2001; Ferrández et al., 1999).

1.4.3 Classification and clustering

A few rule-based coreference systems were built for the MUC-6 and MUC-7 conferences (Appelt et al., 1995; Gaizauskas et al., 1995; Garigliano et al., 1997), but the fact that these conferences conducted large-scale evaluations and developed a considerable amount of annotated data for that purpose contributed to the growth in applying machine-learning methods to the coreference task. The years since then have seen increasing research in coreference resolution systems, and it is to their resolution strategies that I now pay attention. Pronoun resolution systems have continued to be developed though (Yang et al., 2004; Navarretta, 2004; Kehler et al., 2004; Hinrichs et al., 2007), and especially for computational models of dialogue (Strube and Müller, 2003; Frampton et al., 2009). Performance scores are reported in Table 1.3 and will be discussed in 1.4.4.

Two steps Different models have been proposed to partition the mentions of a document into a set of entities on the basis of the linguistic information encoded by the features summarized in 1.4.2. Soon et al.'s (2001) formulation of the task, inspired by the early systems of Aone and Bennett (1995) and McCarthy and Lehnert (1995), has become a standard starting point for anyone attempting to build a coreference system. Under this conception, the coreference task is modeled as a two-step procedure:

1. A *classification* phase that decides whether two mentions corefer or not. It is a binary classification problem in which the probability of mention m_i and mention m_j having a coreferential outcome can be calculated by estimating the probability that:

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

2. A *clustering* phase that converts the set of pairwise classifications into clusters of mentions, creating one cluster for each entity. This phase requires coordinating the possibly contradictory coreference classification decisions from the first phase.

Coreference systems can differ along both dimensions independently. In the classification stage, the coreference probability of two mentions can be predicted by training different machine-learning algorithms such as decision trees (Soon et al., 2001; Ng and Cardie, 2002b; Ng, 2005; Yang and Su, 2007), maximum entropy classifiers (Luo et al., 2004; Ng, 2005; Nicolae and Nicolae, 2006; Ponzetto and Strube, 2006), the RIPPER rule learner (Ng and Cardie, 2002b; Hoste, 2005; Ng, 2005), SVMs (Uryupina, 2007; Rahman and Ng, 2009), memory-based learning (Klenner and Ailloud, 2009; Hoste, 2005), or averaged perceptrons (Bengtson and Roth, 2008; Stoyanov et al., 2009). In the clustering stage, a broad distinction can be drawn between local and global models or, in other words, between mention-pair and entity-based models.

Mention-pair models Mention-pair models can follow different strategies such as *link-first* (Soon et al., 2001; Strube et al., 2002) and *best-first*, the most widely used (Ng and Cardie, 2002b; Yang and Su, 2007; Bengtson and Roth, 2008). The former compares each mention in turn to each preceding mention, from right to left, and the process terminates as soon as the beginning of the text is reached or the classifier returns a coreference probability above 0.5 for a mention pair, in which case the two mentions are clustered into the same entity. In contrast, best-first computes the probability of all mentions preceding the mention under analysis and picks the one with the highest coreference probability (above 0.5), thus making the most confident decision.

First-link and best-link models are mention-pair models that present a major drawback in that they are only locally optimized. Since coreference is a transitive relation,[5] these models simply perform the transitive closure of the pairwise decisions, but do not ensure the global consistency of the entity. For example, *Mr. Clinton* may be correctly coreferred with *Clinton*, but then particular pairwise features may make the model incorrectly believe that *Clinton* is coreferent with a nearby occurrence of *she*, and since the clustering stage is independent from the pairwise classification, the incompatibility between the gender of *Mr. Clinton* and that of *she* will be ignored in building the final cluster. This is what marks the divide between local and global or entity-based models.

Entity-based models Unlike mention-pair models, entity-based ones take advantage of the information provided by other mentions in a preceding, possibly partially-formed, entity. This can be especially helpful when it is difficult to judge whether or not two mentions are coreferent simply from the pair alone, and might provide a mechanism to either recover a missed link or avoid a spurious one. Obviously, transitivity restrictions cannot be enforced by any system that works only on links. To this end, the latest coreference systems work on clusters that allow assessing how well a particular mention matches an entity *as a whole*.

One of the first systems to move in this direction was Luo et al. (2004), who consider all clustering possibilities (i.e., entity partitions) by searching in a Bell

[5] By the transitivity property, it follows that if mention a is coreferent with b and b is coreferent with c, then a is coreferent with c.

tree representation, and cast the coreference resolution problem as finding the best path from the root node to the leaves (where each leaf is a possible partition). The different partition hypotheses are built by using either a standard mention-pair classifier or an entity-mention one, which determines the probability that a mention refers to a given entity. Surprisingly enough, however, the latter underperforms the former. Nicolae and Nicolae (2006) argue that "despite the fact that the Bell tree is a complete representation of the search space, the search in it is optimized for size and time, while potentially losing optimal solutions." In addition, Luo et al. (2004) allude to the lower number of features (twenty times less) used by the entity-mention model as a possible reason for the drop in performance.

Since Luo et al.'s (2004) proposal for a global search in a Bell tree, other ways of globally optimizing the clustering decision have been suggested: a first-order probabilistic model that allows features based on first-order logic over a set of mentions (Culotta et al., 2007); integer linear programming to enforce the transitivity constraint (Finkel and Manning, 2008; Klenner and Ailloud, 2009); a graph-cut algorithm on a graph representation where the nodes represent mentions and the edges are weighed by the pairwise coreference probability (Nicolae and Nicolae, 2006; Ng, 2009); a conditionally-trained graph model (McCallum and Wellner, 2005; Wick and McCallum, 2009); an online learning model that learns the optimal search strategy itself (Daumé III and Marcu, 2005); or inductive logic programming to learn from the relational knowledge of a mention, an entity, and the mentions in the entity with a set of first-order rules (Yang et al., 2008). Although these models ensure global consistency, not all of them include cluster-level features (McCallum and Wellner, 2005). The design of appropriate cluster-level features has been little explored.

Unsupervised models It is among globally optimized systems that we find the few unsupervised systems that exist for coreference resolution: Haghighi and Klein (2007) employ a non-parametric Bayesian generative model based on a hierarchical Dirichlet process, and Poon and Domingos (2008) introduce mention relations like apposition and nominal predicates by developing an unsupervised learning algorithm for Markov logic. Both Haghighi and Klein (2007) and Poon and Domingos (2008) impose a prior on the number of clusters, which is not the case with Ng's (2008) generative system. Ng (2008) modifies the expectation-maximization (EM) algorithm so that the number of clusters does not have to be predetermined, and redefines the E-step to calculate the *n* most probable coreference partitions using a Bell tree (Luo et al., 2004). Haghighi and Klein's (2010) generative, unsupervised system is meant to address semantic compatibility between headwords by exploiting a large inventory of distributional entity types. Finally, Cardie and Wagstaff's (1999) early system lies between supervised and unsupervised learning. It applies clustering on feature vectors that represent mentions with the aim of creating one cluster for each entity, but it is not fully unsupervised because the distance metric used for comparison uses fixed weights that are heuristically decided.

Ranking models The strategy of ranking can be considered an intermediate step between local and global models. A ranker allows more than one candidate mention to be examined simultaneously and, by determining which candidate is most probable, it directly captures the competition among them. The first ranking model by Connolly et al. (1994), which ranks two (a positive and a negative) candidate mentions at a time, is used by Yang et al. (2003) under the name of *twin-candidate model* and by Iida et al. (2003) under the name of *tournament model* for Japanese zero anaphora. The candidates, however, are compared in a pairwise fashion. In contrast, Denis and Baldridge's (2008) ranker considers the entire candidate set at once. Ng (2005) makes a different use of ranking and recasts the coreference task as ranking candidate partitions generated by different mention-pair systems. Thus he can benefit from the strengths of different methods. It is not a really global approach, however, as the candidate partitions are all generated by mention-pair models. Finally, Rahman and Ng (2009) propose a cluster-ranking approach that combines the strengths of mention rankers and entity-mention models.

Enhancing strategies Other methods that have been used for enhancing coreference resolution are the separation of resolution modules for pronouns, proper nouns, and full NPs (Morton, 2000; Müller et al., 2002; Hoste, 2005; Ng, 2005; Haghighi and Klein, 2007; Luo, 2007; Denis and Baldridge, 2008); and explicitly determining the probability of a mention to be discourse-new, either as a separate classifier in a cascade setup (Ng and Cardie, 2002a; Bean and Riloff, 1999; Vieira and Poesio, 2000; Uryupina, 2003; Yang et al., 2003; Kabadjov, 2007; GuoDong and Fang, 2009) or by coordinating discourse-new and coreference probabilities together (Luo, 2007; Ng, 2009), including the joint inference or learning of discourse-new detection and coreference resolution (Denis and Baldridge, 2007; Poon and Domingos, 2008; Rahman and Ng, 2009).

The learning process can also be boosted by filtering out expletive pronouns (Evans, 2000; Boyd et al., 2005; Bergsma et al., 2008) and non-referential indefinite NPs (Byron and Gegg-Harrison, 2004). For Danish, Navarretta (2009a) develops an automatic classifier of neuter pronouns and demonstrative pronouns into a range of functions including non-referential, cataphoric, deictic, anaphoric with an NP antecedent, anaphoric with a clausal or sentential antecedent, and vague (i.e., the antecedent is implicit in discourse).

I will now focus on the performance scores of the different learning-based systems discussed so far and present the problem of evaluation. Notice that Table 1.3 in the next section also includes Haghighi and Klein's (2009) system, which is an isolated case of rule-based approach in recent years. It achieves performance comparable to state-of-the-art learning-based systems despite using only a few syntactic and semantic constraints (e.g., head match, agreement, apposition). It clearly contrasts with some of the complex learning algorithms that have been implemented.

1.4.4 Evaluation

Like other NLP tasks, evaluating a coreference resolution system includes not only assessing its performance, but also weighing its overall benefits compared to the state of the art. Quantifying how well a system performs is not straightforward. Byron (2001) and Mitkov and Hallett (2007) bring attention to inconsistencies when reporting results in the sister task of pronoun resolution. They point out that the sorts of pronouns in scope vary between studies and that most algorithms benefit from post-edited outputs, both of which have an obvious effect on standard precision and recall. Similar problems—and even worse—are encountered in the evaluation of coreference systems.

True and system mentions In coreference resolution, a major difficulty for defining an appropriate performance metric arises from not knowing the total number of entities beforehand. This is aggravated by the fact that the mentions resolved by the system (*system mentions*) might not directly map onto the mentions of the gold standard (*true mentions*) if they are automatically detected. Moreover, the different mentions considered by different annotation schemes (e.g., only mentions in multi-mention entities by MUC, only mentions of specific semantic types by ACE) have a direct effect on the complexity of solving a specific text. As a result, performances tend to vary considerably across different corpora (Stoyanov et al., 2009). This is why scores are separated according to the test data in Table 1.3.

Current scoring metrics Just as the MUC program can be considered as the starting point of coreference resolution systems and of large-scale coreferentially annotated data, it was also the first to define a scoring metric, known as the MUC metric (Vilain et al., 1995). Despite its wide use, numerous weaknesses of this metric have been pointed out on several occasions (Bagga and Baldwin, 1998; Ng, 2005; Nicolae and Nicolae, 2006; Denis and Baldridge, 2008; Finkel and Manning, 2008) and alternative metrics have been proposed, among which B^3 (Bagga and Baldwin, 1998) and CEAF (Luo, 2005) remain the most widely used alternatives. These measures are discussed in more detail in Chapter 5, but a brief summary of their formulas follows:[6]

- MUC-metric

$$\text{R} = \frac{\text{\# common links in true and system partition}}{\text{\# minimum links for true partition}}$$

$$\text{P} = \frac{\text{\# common links in true and system partition}}{\text{\# minimum links for system partition}}$$

[6] Each metric is computed in terms of recall (R), a measure of completeness, and precision (P), a measure of exactness. The F-score corresponds to the harmonic mean: F-score = 2 · P · R / (P + R).

- B^3

$$\mathrm{R} = \frac{\sum_{i=1}^{n} \frac{\text{\# common mentions in true and system entity of mention}_i}{\text{\# mentions in true entity of mention}_i}}{n}$$

$$\mathrm{P} = \frac{\sum_{i=1}^{n} \frac{\text{\# common mentions in true and system entity of mention}_i}{\text{\# mentions in system entity of mention}_i}}{n}$$

- CEAF-ϕ_3

$$\mathrm{R/P} = \frac{\text{\# common mentions in best one-to-one aligned true and system entities}}{\text{\# mentions in true/system partition}}$$

The B^3 metric was designed to address the two main shortcomings presented by MUC, namely its preference for systems that create large entities and its ignorance of correctly clustered singletons. The CEAF metric was in turn proposed to solve a weakness found in B^3, namely that an entity can be used more than once when aligning true and system entities. The acknowledged drawbacks notwithstanding, the MUC metric has continued to be used for two main reasons: (i) for comparison purposes with the oldest systems that only report MUC scores, and (ii) due to the lack of agreement on a standard metric as none has been shown to be clearly superior. Consequently, the evaluation of coreference systems is currently done by providing either one or two or all three scoring metrics, as displayed in Table 1.3.

State-of-the-art scores The MUC metric is the only one for which we have the scores of almost all systems, but this is rather useless since making quality judgments based on a faulty metric (see the arguments in Chapter 5) would be clearly misleading. Denis and Baldridge (2008) strongly advocate that coreference results should never be presented in terms of MUC scores alone. Further evidence that it is not possible to rely on the MUC metric alone comes from the different system rankings obtained by MUC and B^3 (compare, for instance, the antepenultimate and the previous row in Table 1.3). The problem is that some systems report results using either B^3 or CEAF, and these are clearly not commensurate.

All in all, it is only reasonable to conclude that there is no clear winner for the state of the art. The lack of a reliable metric, the use of different corpora (and of different portions of the same corpus) and the reliance on true or system mention boundaries (e.g., unlike Harabagiu et al. [2001], Soon et al. [2001] and Ng and Cardie [2002b] assume no preprocessing at all of the data sets) make any comparison between different systems meaningless. Conducting a qualitative evaluation, on the other hand, is only possible for the few systems that have been released,[7] and their output appears to link a considerable number of non-coreferent phrases as coreferent, and vice versa.

[7] Publicly available coreference systems include OpenNLP (http://opennlp.sourceforge.net), BART (Versley et al., 2008), Reconcile (Stoyanov et al., 2010), the Illinois Coreference Package (Bengtson and Roth, 2008), CoRTex (Denis and Baldridge, 2008), CherryPicker (Rahman and Ng, 2009), and ARKref (http://www.ark.cs.cmu.edu/ARKref).

System	Mentions	MUC			B^3			CEAF		
		P	R	F	P	R	F	P	R	F
MUC-6 data										
Cardie and Wagstaff (1999)	System	54.6	52.7	53.6						
Morton (2000)	System	79.6	44.5	57.1						
Harabagiu et al. (2001)	True	92	73.9	81.9						
Soon et al. (2001)	System	67.3	58.6	62.6						
Ng and Cardie (2002b)	System	78.0	64.2	70.4						
Yang et al. (2003)	True	80.5	64.0	71.3						
Luo et al. (2004)	True			85.7						76.8
McCallum and Wellner (2005)	True	80.5	64.0	71.3						
Choi and Cardie (2007)	System	69.3	70.5	69.9						
Haghighi and Klein (2007)	True	80.4	62.4	70.3						
Finkel and Manning (2008)	True	89.7	55.1	68.3	90.9	49.7	64.3			
Poon and Domingos (2008)	True	83.0	75.8	79.2						
Haghighi and Klein (2009)	True	87.2	77.3	81.9	84.7	67.3	75.0	72.0	72.0	72.0
Stoyanov et al. (2009)	System			68.5			70.9			
MUC-7 data										
Soon et al. (2001)	System	65.5	56.1	60.4						
Ng and Cardie (2002b)	System	70.8	57.4	63.4						
Yang et al. (2003)	True	75.4	50.1	60.2						
Uryupina (2007)	System	67.0	50.5	65.4						
Stoyanov et al. (2009)	System			62.8			65.9			
ACE-2 data										
Luo et al. (2004)	True									73.1
Nicolae and Nicolae (2006)	True	91.1	88.2	89.6				82.7	82.7	82.7
	System	52.0	82.4	63.8						41.2
Denis and Baldridge (2007)	True	77.1	63.6	69.7						
Yang and Su (2007)	System	73.9	56.5	64.0						
Denis and Baldridge (2008)	True	75.7	67.9	71.6	79.8	66.8	72.7	67.0	67.0	67.0
Finkel and Manning (2008)	True	83.3	52.7	64.1	90.2	62.6	73.8			
Poon and Domingos (2008)	True	68.4	68.5	68.4	71.7	66.9	69.2			
Ng (2009)	System	69.2	55.0	61.3				59.6	63.7	61.6
Stoyanov et al. (2009)	System			66.0			78.3			
ACE-2003 data										
Ponzetto and Strube (2006)	System	84.2	61.0	70.7						
Ng (2008)	True	69.9	51.6	59.3				61.1	61.1	61.1
	System	63.3	48.8	54.7				55.6	60.0	57.7
Yang et al. (2008)	System	60.5	63.4	61.8						
Stoyanov et al. (2009)	System			67.9			79.4			
ACE-2004 data										
Luo and Zitouni (2005)	True							82.0	82.0	82.0
Culotta et al. (2007)	True				86.7	73.2	79.3			
Haghighi and Klein (2007)	True	65.0	61.8	63.3						
Bengtson and Roth (2008)[a]	True	82.7	69.9	75.8	88.3	74.5	80.8			
Poon and Domingos (2008)	True	69.1	69.2	69.1						
Haghighi and Klein (2009)[a]	True	74.8	77.7	79.6	79.6	78.5	79.0	73.3	73.3	73.3
	System	67.5	61.6	64.4	77.4	69.4	73.2			
Stoyanov et al. (2009)[a]	System			62.0			76.5			
Wick and McCallum (2009)	True	78.1	63.7	70.1	87.9	76.0	81.5			
Haghighi and Klein (2010)[a]	System	67.4	66.6	67.0	81.2	73.3	77.0			
ACE-2005 data										
Luo (2007)	True							84.8	84.8	84.8
Rahman and Ng (2009)	True	83.3	69.9	76.0	74.6	56.0	64.0	63.3	63.3	63.3
	System	75.4	64.1	69.3	70.5	54.4	61.4	62.6	56.7	59.5
Stoyanov et al. (2009)	System			67.4			73.7			
Haghighi and Klein (2010)[b]	System	74.6	62.7	68.1	83.2	68.4	75.1			
Haghighi and Klein (2010)[c]	System	77.0	66.9	71.6	55.4	74.8	63.8			

[a] ACE-2004 test set utilized in Culotta et al. (2007).

[b] ACE-2005 test set utilized in Stoyanov et al. (2009).

[c] ACE-2005 test set utilized in Rahman and Ng (2009).

Table 1.3: Summary of coreference resolution system performances

1.5 Connecting thread

The four aspects of coreference covered by this thesis—theory, annotation, resolution, and evaluation—are closely interrelated yet separable. For space and clarity reasons, I limit myself to mainly one facet of the problem in each paper, but putting it always into perspective as to how it affects and connects with the rest of the problem. In this section, the different perspectives are integrated. I discuss the overall framework, the findings across the papers that show signs of disruption and the new directions that I have taken to meet the needs of the coreference resolution task.

1.5.1 Methodological framework

From the outset, I took a corpus-based approach to the coreference problem, setting the use of real data as a priority. My main concern was to study coreference as it occurs in natural data. Consequently, many problems posed in this thesis fall outside the focus of theoretical linguistics and psycholinguistics, where analyses tend to be confined to carefully constructed or selected examples. Because they are rarely longer than two sentences, such examples lack the particular relations that are only possible in the context of a long written discourse. More precisely, I focused my research on newspaper texts. A total of six corpora were used:

AnCora (Recasens and Martí, 2010) 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 (developed as part of this thesis work).

ACE (Doddington et al., 2004) The set of English data for the ACE 2003, 2004 and 2005 programs includes newswire, newspaper and broadcast news from the TDT collection. They are annotated with ACE entity types (e.g., person, organization, location, facility, geo-political entity, etc.), entity subtypes, mention class (e.g., specific, generic, attributive, etc.), and mentions of the same entity are grouped together. The corpora were created and are distributed by the Linguistic Data Consortium.

OntoNotes (Pradhan et al., 2007a) The English OntoNotes Release 2.0 corpus 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, word senses (linked to an ontology), and coreference information. The OntoNotes corpus is distributed by the Linguistic Data Consortium.

KNACK-2002 (Hoste and De Pauw, 2006) A Dutch corpus containing 267 documents from the Flemish weekly magazine Knack. They are manually annotated with coreference information on top of semi-automatically annotated PoS tags, phrase chunks, and NEs.

TüBa-D/Z (Hinrichs et al., 2005) A German 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 in order to obtain a corpus license.

LiveMemories (Rodríguez et al., 2010) An Italian corpus under construction that 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.

In line with functional linguistics (Halliday and Hasan, 1976; Gundel et al., 1993), I adopted a discourse representation approach to reference, locating the coreference phenomenon within the discourse model projected by language users, and replacing the notion of "world referents" by "discourse referents" (Karttunen, 1976; Webber, 1979; Kamp, 1981). The discourse model is built up dynamically and is continuously updated, including not only the entities explicitly mentioned, but also those that can be inferred from them. In the computational field, the view that entities belong to the real world has dominated (Ng, 2009; Finkel and Manning, 2008), while only a minority have opted for the discourse model hypothesis (Poesio, 2004a; Bengtson and Roth, 2008; Denis and Baldridge, 2009). Sticking to the real world might seem to avoid unnecessary theoretical jargon, but it runs quickly into a host of conceptual problems, starting with fictional and hypothetical entities: What real-world entity does *Superman* point to? Or what real-world entity do you refer to when making plans about your next car? In brief then, in the definition of coreference given by van Deemter and Kibble (2000),

> NP_1 and NP_2 *corefer* if and only if $\text{Referent}(\text{NP}_1) = \text{Referent}(\text{NP}_2)$,
> where Referent(NP) is short for "the entity referred to by NP"

I further specify Referent(NP) as "the discourse entity referred to by NP in the discourse model."

That said, my guiding principle was to achieve a good compromise between linguistic accuracy and computational possibilities, adopting operational definitions whenever possible. In order to make the scope more manageable, some limitations were necessary. This thesis focuses on intra-document coreference,[8] including identity-of-reference anaphora but excluding strictly anaphoric phenomena (Hirst, 1981) such as identity-of-sense anaphora (3), ellipsis (4), bound anaphora (5), and bridging anaphora (6), which requires the reader to draw an inference to identify the textual anchor through a relation other than identity.

(3) Lyle drove [a car]. Maria drove [one], too.

[8] Given its discourse function, genuine coreference occurs within a single discourse unit, or across a collection of documents linked by topic. This work views *cross-document* coreference as an NLP application which assumes that there is an underlying global discourse that enables various documents to be treated as a single macro-document.

(4) George was bought a huge box of [chocolates] but few [ø] were left by the end of the day.

(5) [Every TV network] reported [its] profits.

(6) I looked into [the room]. [The ceiling] was very high.

Similarly, attributive (7), predicative (8) and appositive (9) relations fall outside the scope of the current work for not being referential relations, thus following annotation schemes like MATE and OntoNotes that distinguish between coreference and predicative links (Poesio, 2004b; Pradhan et al., 2007b).

(7) The [Eyjafjallajökull] volcano, one of Iceland's largest, had been dormant for nearly two centuries.

(8) The Eyjafjallajökull volcano is [one of Iceland's largest].

(9) The Eyjafjallajökull volcano, [one of Iceland's largest], had been dormant for nearly two centuries.

A second limitation was to concentrate on referential acts performed by NPs. By NPs I include pronouns, proper nouns, and phrases headed by a common noun, which correspond to ACE "pronominals," "names," and "nominals," respectively. The terms "pronouns" and "full NPs" are used to distinguish the two last groups from the first one. Non-nominal expressions, however, were not completely excluded, as verbs, clauses, and discourse segments (Webber's [1979] *discourse deixis*) were also included in the annotation of AnCora (see preliminary work in Recasens [2008]).

From a computational perspective, I chose to investigate machine learning techniques as they offer great potential to discover patterns and general tendencies not discernible to the human eye, and coreference resolution had been a target of learning-based approaches since the mid-90s. Memory-based learning was applied using TiMBL v.6.1.0. (Daelemans and Bosch, 2005) after testing several other machine learning packages like maximum entropy (Berger et al., 1996) and decision tree models (Quinlan, 1993). The preference for TiMBL was based mainly on its robustness to sparse feature spaces and its user-friendly properties like the display of the feature ranking.

For the evaluation, I followed a twofold strategy: (i) the most widely accepted measures (MUC-metric, B^3 and CEAF) were used to *quantitatively* assess system performance, and (ii) the accuracy of the system outputs was also evaluated *qualitatively* by manually performing an error analysis of a sample of automatically annotated texts.

1.5.2 Signs of disruption

During the course of this research, I accumulated a series of findings that suggested there were root problems in the coreference task. Thus, in a change of direction, my attention was diverted from refining the resolution stage to reconsidering the coreference problem from the start toward formulating a workable solution. I highlight here the major signs of disruption that were noticed, without entering into the technical details that can be found in the papers that follow.

Degrees of referentiality Annotating a corpus forces one to consider every single relation in the data instead of simply selecting the easy, clear-cut relations. In the present case, given that I aimed to mark coreference relations between NPs (Chapter 2), a need arose early to distinguish not only attributive (7), predicative (8) and appositive (9) phrases, but to further separate all referential NPs from non-referential ones (10).

(10) The Eyjafjallajökull volcano, one of Iceland's largest, had been dormant for [nearly two centuries] before returning gently to [life] in the late evening of March 20, 2010, noticeable at [first] by the emergence of a red cloud glowing above the vast glacier that covers it. In the following days, fire fountains jetted from a dozen vents on the volcano, reaching as high as [100 meters].

Although it is possible to identify certain classes of non-referentiality like duration phrases (e.g., *nearly two centuries*), measure phrases (e.g., *100 meters*) and idioms (e.g., *at first*), the referential/non-referential distinction tends to blur at the margins. As an example, consider *life* in (10). It is on the border of becoming grammaticalized, hence the lack of determiner, similar to phrases like *go to school* or *go home*. Such cases support Fraurud's (1992) idea that being a discourse referent is not a matter of either-or. Contemplating degrees of individuation, i.e., different levels of representation in the discourse model, seems to be more in accordance with natural data.

Overlooked singletons Given that coreference is a binary relation, it is common for annotation efforts to mark only multi-mention entities. This is not problematic if referentiality is also marked, but it is otherwise, as then all non-coreferent NPs are counted as singletons by default—there being no other way to extract singletons from the manual annotation—and this is obviously at the cost of having non-referential NPs introduce considerable noise among mentions. I encountered such a difficulty when comparing OntoNotes and AnCora (Chapter 4) as well as when extracting the SemEval task datasets from OntoNotes, KNACK-2002 and TüBa-D/Z (Chapter 6).

In retrospect, the issue of singletons has more implications than initially thought. Detecting non-coreference is as important as detecting coreference and this is why several coreference systems model so-called (but wrongly called) "anaphoricity" to avoid treating a mention as a subsequent mention when it is not (Ng, 2004; Luo, 2007; Denis and Baldridge, 2007). Mentions classified as non-anaphoric, however, will still be considered again and again as first mentions throughout the resolution process. Rather, the preponderance of isolated mentions (60% of all NPs, Table 4.1; 53% of all mentions, Table 2.3) suggests that what would be of great help for a coreference system is an automatic classifier of singletons that filters out mentions that do not need to be considered as either subsequent or first mentions. I did some preliminary experiments in this direction, but no linguistic properties of NPs seem to be distinctive of isolated mentions.

On the evaluation side, the large number of singletons accounts for the all-singletons baseline (i.e., a system outputting every mention as an entity) setting so high a baseline (Chapters 4 and 6), especially for corpora that are not restricted to any entity types, like AnCora, OntoNotes, TüBa-D/Z and LiveMemories. The difference in the number of singletons between OntoNotes and ACE is the reason why resolution systems score higher on the former than on the latter according to class-based measures like B^3 and CEAF. Stoyanov et al. (2009) come to a similar conclusion with respect to the reason why ACE systems obtain higher B^3 scores than MUC systems. The major source of dissatisfaction with current metrics is precisely the treatment of singletons.

Entity distribution The focus on the key role of coreference chains in discourse cohesion and coherence has naturally led to the assumption that a good number of entities in a discourse are referred to multiple times. I have already argued against this on the basis of singletons. But there is another remark to be made. The majority of multi-mention entities are mentioned not many but just a few times. It emerges from the annotated data that the average size per entity ranges from three to four mentions, being two mentions the most frequent size (Table 2.5). In sum, the overall picture is that the distribution of entities in a discourse is skewed with a majority of singletons playing a peripheral role, and a second large group of entities that are mentioned a couple of times, leaving the number of entities that are mentioned more than twice at approximately two per document.

This brings to the fore the split between short-term and long-term referents made by Karttunen (1976), or between local-focus and global-focus entities by Grosz and Sidner (1986). There have been very few attempts—Daumé III and Marcu's (2005) measure of "decayed density" is one—to make such a distinction in coreference systems, but it would probably prove useful. In order for a system to be able to decide on the centrality of an entity, a strategy that goes beyond pairwise learning features needs to be implemented.

Borderline coreference A key observation is that not all entity types have the same coreference potential. People and organizations (typically introduced by proper nouns) tend to be more coreferred than locations or dates, for instance. This supports the claim that the ontological type of entities makes a difference in terms of individuation (Fraurud, 1992) and so in coreference. And it is precisely among the highest individuated types, namely people and organizations, that most inter-coder disagreements occur. Examples like (11) and (12) are hard to classify as either coreferent or non-coreferent.

(11) For centuries here, [the people] have had almost a mystical relationship with Popo, believing the volcano is a god. Tonight, [they] fear it will turn vengeful.

(12) According to Aznar, Parliamentary Monarchy "is not only the expression of [the modern Spain], but it is also a symbol of stability and permanence" . . . According to Aznar, the Crown can "guarantee and express" that [Spain] can have "more ambitions, intentions, and goals."

These borderline relations are not a marginal phenomenon. I had to deal with them not only during the annotation of the AnCora corpus, but also when training the proposed system on the ACE and OntoNotes corpora. The troublesome nature of these examples became definitely apparent when the same texts annotated by both ACE (13-a) and OntoNotes (13-b) were compared (square brackets identify the mentions annotated as coreferent).

(13) 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 different annotation is clear evidence—as it is the low or *just* reasonable level of inter-annotator agreement reported by previous efforts (Poesio and Vieira, 1998; Müller, 2007; Poesio and Artstein, 2005)—of a severe weakness in the definition of coreference. The prevailing definition is too general to account for naturally occurring data at large. It makes no mention of metonymy, which abounds in natural language, nor of underspecified entities, which I briefly touched upon in connection with abstract objects (Recasens, 2008). The picture that emerges is of a much more complicated phenomenon.

No universal rules It was only gradually that I strengthened my understanding of the complex picture of coreference, simultaneously as machine learning experiments were carried out. These in turn helped lay bare the shortcomings of the coreference resolution task and revealed the limitations of machine learning to address the problem.

The success that machine learning has seen in NLP tasks such as PoS tagging or syntactic parsing owes much to the fact that their outputs can be inferred from surface properties such as the surrounding context (*n*-grams) or the distribution of tags and words in some relevant contexts. Following a similar path, coreference resolution has tried to exploit morphological, syntactic and semantic formal cues. But it has not achieved the same satisfactory results as PoS tagging or syntactic parsing. This failure can be attributed to three main causes (Chapter 3). First, there appear to be very few rules that systematically govern coreference relations (Hoste, 2005; Denis, 2007). This is also demonstrated by the plateau obtained with increased amounts of training data (Chapter 4). Second, the learning features that have been used so far fail to capture pragmatic cues that are needed to detect specific, unique, coreference relations. And new effects are still being discovered (Arnold and Griffin, 2007). Third, a complex set of yet uncontrolled interactions between features are at play (Luo and Zitouni, 2005), which explains the *just* moderate improvement in performance achieved by Uryupina's (2008) 351 features versus Soon et al.'s (2001) 12 features. This also explains the importance of feature selection (Ng and Cardie, 2002b; Hoste, 2005). All these prevent learning algorithms from evolving a general, universal enough, model of coreference.

Psycholinguistic and cognitive studies for English (Arnold et al., 2000; Gordon et al., 1993; Crawley et al., 1990; Stevenson et al., 1994; Kehler et al., 2008), Spanish (Carreiras and Gernsbacher, 1992) and Catalan (Mayol and Clark, 2010) have provided empirical evidence for many of the features used by coreference resolution systems in the last fifteen years. To the extent that experiments have demonstrated that these features prompt readers to favor certain pronoun interpretations over others, it has been assumed that they should equally work in automatic systems, either in the form of constraints and preferences or in the form of learning features. Nevertheless, an important reason why they have not been as successful as expected is the gap between laboratory or focus-oriented linguistic studies and real data, where all the different phenomena occur and interact at once. This is not to deny the value and importance of these studies, but to evaluate their contribution to the task at hand. As Krahmer (2010) points out, psycholinguists and computational linguists have different goals. Whereas the former are interested in showing a general effect and learning about human memory, the latter are interested in achieving overall good performance and so care about the points in the dataset that are processed incorrectly by their model.

Broader theories of reference such as Accessibility Theory (Ariel, 1988), Givenness Hierarchy (Gundel et al., 1993) or Centering Theory (Grosz et al., 1995) provide interesting criteria according to which referring expressions are arranged along a scale. Nevertheless, the multiple factors involved in assessing the degree of "accessibility" or "cognitive status" explain why such notions are so hard to capture computationally and run into implementation problems (Poesio et al., 2004b). Distance, for instance, is a crucial factor determining degree of accessibility, but it is not the only one (Ariel, 2001). In a similar vein, Tetreault (1999) discusses inconsistent uses of gender, Barbu et al.'s (2002) corpus-based study finds that almost a quarter of plural pronouns corefer with a constituent other than a plural NP, and Poesio et al. (2004b) point out that entity coherence between utterances is much less strong than expected.

The examples below are meant to illustrate some of the problems that coreference resolution systems should be—but are not—able to cope with. Real data includes numerous counter-examples to features such as number agreement (14), definiteness as an accessibility marker (15), indefiniteness as a marker of new information (16), and even head match (17) (18).

(14) a. [Madeleine Albright] meets tomorrow with [Ehud Barak] and [Palestinian Authority President Yasser Arafat]. [They]'re expected to meet in the afternoon.

b. Madeleine Albright meets tomorrow with [Ehud Barak] and [Palestinian Authority President Yasser Arafat]. [They]'re expected to meet separately with Albright.

(15) a. [The Eyjafjallajökull volcano, one of Iceland's largest,] had been dormant for nearly two centuries . . . fire fountains jetted from a dozen vents on [the volcano], reaching as high as 100 meters.

b. [The Eyjafjallajökull volcano, one of Iceland's largest,] had been dormant for nearly two centuries.

(16) a. [A new study detailing the uncompensated work burden on family doctors] points to the need to change how they are paid.

b. [Postville] might be catching up with the rest of America . . . The plants helped spur economic development in [a town that had long been stagnant].

(17) a. ABC's Gillian Finley begins in [Palestinian Gaza]. In [Gaza] today, Israeli soldiers opened fire on school boys, throwing stones.

b. President Clinton was in [Northern Ireland] when he heard the Supreme Court decision . . . Clinton thanked the government of [Ireland] for accepting two prisoners.

(18) a. [A hundred artists] will participate in the event . . . Of [the numerous artists willing to participate in this celebration], half will do it at the beginning, and the other half at the end of the celebration.

b. [A hundred artists] will participate in the event . . . It has not been possible to count with [all the numerous artists willing to participate in this celebration] due to time limitations.

Pragmatics Since learning algorithms are not provided with any explicit information about pragmatics or world knowledge, it is very hard for them to discriminate between (14-a) and (14-b). Similarly, in the case of definite NPs—the most frequent form of NP in Spanish and Catalan—although they are typically considered to refer back to a previously introduced entity (15-a), over 50% of the time they are the form of isolated or first mentions (15-b) (Fraurud, 1990; Poesio and Vieira, 1998; Recasens et al., 2009b; Recasens, 2009). Indefinite NPs are assumed to accomplish the opposite function, i.e., to mention an unknown entity for the first time (16-a), but again this rule is not without exceptions (16-b). As for proper names, they can equally introduce an NP as corefer with a previously introduced entity (Table 2.3). In fact, all theories that arrange referring expressions on a scale (Ariel, 1988; Gundel et al., 1993) agree that additional, pragmatic factors can override the principles they propose. To add but one more example, Hervás and Finlayson (2010) show that 18% of referring expressions in news and narrative are descriptive, i.e., they provide additional information not required for distinction. This can be contrary to the principle that considers long definite descriptions to be low accessibility markers (Ariel, 1988).

Within this context of non-generalizable features, head match appears as the most robust feature. Of them all, it is clearly the feature that solves the most relations with the least error, although it does not work all the time either (Vieira and Poesio, 2000; Soon et al., 2001; Uryupina, 2008; Elsner and Charniak, 2010), as exemplified by the positive cases (17-a) (18-a) versus the negative ones (17-b) (18-b). Nevertheless, solving coreference relations involving proper nouns or full NPs that do not have the same head remains as one of the hardest problems (Haghighi and Klein, 2010). Improving recall at the minimal cost of precision seems to be only surmountable up to 80% F-score. The remaining 20% is actually very harmful for the linguistic quality of the results, although little attention has been paid to this issue. I return to this point below. Surely the features that have been proposed are all important in some aspect, but we lack something: The way different knowledge sources and different features contribute and interact together escapes the current methods.

Preprocessing effects Another expectation that was not fully met concerns the use of automatic preprocessing information versus perfect morphological and syntactic information. In order to determine the extent to which performance drops when gold-standard information is not available, the same system was run on the same texts varying the source of preprocessing information (Chapter 4). To my surprise, the performance drop was not very pronounced. I found two explanations. First, relevant features like head match are unaffected by the quality of preprocessing. Secondly, the learning algorithm reranks features in such a way that morphological information gains position over syntactic information, thus downplaying the noise brought by automatic preprocessing tools. It appears to be the case that, with a good set of basic shallow features, learning can do almost as well without rich features that depend on deep parsing.

The consequences of using automatic preprocessing are more severely felt in the detection of mention boundaries (Stoyanov et al., 2009), which is identified by Uryupina (2008) as a main cause of spurious links. Using automatically-predicted mention boundaries causes drastic drops in performance as shown in Table 1.3 if the scores of systems that are tested on both true and system mentions are compared (Nicolae and Nicolae, 2006; Ng, 2008; Haghighi and Klein, 2009; Rahman and Ng, 2009). This was the cause of the design error in the definition of the gold and regular evaluation settings of the SemEval task (Chapter 6): giving true mention boundaries in only the former setting prevented us from being able to draw conclusions about performance improvements thanks to the use of gold preprocessing. Contrary to Stoyanov et al. (2009), who argue that the experimental setting with gold mention boundaries is "rather unrealistic" and makes the coreference task substantially easier, I believe that the problem of mention detection is of a different order. It pertains to syntax in the first place, and to referentiality detection in the second place. Thus, evaluating the tasks of coreference resolution and mention detection as a single task is not only confusing but makes different performances incommensurate.

The head-match and all-singletons baselines More and more features as well as newer and more sophisticated resolution models have been suggested in recent years, but one is struck when their results are compared with those obtained by two naïve baselines (Chapter 4): (i) linking all mentions that share the same head ("the head-match baseline"), and (ii) not predicting any coreferent mention but only singletons ("the all-singletons baseline"). The superiority of head match is supported by the fact that it forms the basis of Stoyanov et al.'s (2009) coreference performance prediction measure that, given a dataset, predicts its resolution complexity. Additionally, Markert and Nissim (2005) identify the large number of definite NPs that are covered by simple head matching as one of the problems in comparing algorithm results for coreference.

What is worrying is that these baselines, especially the all-singletons one, are hardly ever reported in coreference resolution papers. Cardie and Wagstaff (1999) do give the performance of head match on the MUC-6 test data, admitting that it "performs better one might expect." This baseline is also given for the MUC-7 test data by Soon et al. (2001) and Uryupina (2007). The 5-percentage-point difference between their respective baselines is explained by the use of system mentions. Thus, they test on different sets of mentions. Soon et al.'s (2001) system only outperforms head match by 5%, while Uryupina's (2007) outperforms her baseline by 15%.

Depending on the corpus, the all-singletons baseline alone can achieve scores as high as 84% B^3 and 73% CEAF (for OntoNotes). The same baseline falls down to 67% B^3 and 50% CEAF for ACE due to the smaller number of singletons. When adding head match, the ACE scores go up to 76% B^3 and 66% CEAF (note the large number of proper nouns in the ACE corpora). In contrast, overmerging mentions results in a higher MUC score for ACE (68%) than for OntoNotes (56%). The significance of these figures lies in that state-of-the-art systems like the one by Luo et al. (2004) obtain 77% B^3, 73% CEAF, and 81% MUC (on the ACE-2 dataset), which are not so far from these naïve baselines, especially if the enormous effort invested in those systems is taken into account.

Uninformative scores Shortcomings in the way scoring measures are computed lead to high scores of simple baselines like all-singletons or the opposite, i.e., overmerging. With the MUC score, for instance, clustering a mention into the wrong entity is penalized twice as much as merging two gold entities (Poesio et al., forthcoming), thus its bias toward overmerging. Not only MUC but also B^3 and CEAF are biased toward different types of output (Chapters 4 and 5). Consequently, the system's final score turns out to depend more on the corpus characteristics than on the resolution strategy. MUC will rank highest the system outputting the largest number of links, and B^3 will reward high-recall systems that are good at detecting singletons and at linking mentions with the same head—even if linking non-coreferent same-head mentions (Elsner and Charniak, 2010). CEAF, while reaching a better compromise, is still strongly influenced by large numbers of singletons. The disagreement in ranking different systems between the three measures makes them useless for their purpose, namely evaluating coreference resolution models. Thus, it was not possible to draw definite conclusions about the SemEval shared task (Chapter 6), as each measure ranked the participating systems in a different order (Table 6.5).

The quality problem just worsens as the practice of reporting *only* performance scores—and no sample output—becomes more common. At least, Soon et al. (2001) and Uryupina (2008) perform an error analysis. Experience reveals that by looking at the actual output, we might realize that mentions such as *the new president* and *the old president* are linked because of sheer head match. Such an error is still very common in state-of-the-art systems (Haghighi and Klein, 2010). The quality of state-of-the-art coreference systems is far from satisfactory, and the current evaluation metrics, instead of being helpful in this respect, contribute to obscuring the evaluation. Obviously, a qualitative analysis presupposes an understanding of the task definition, and this brings us back to some of the earlier points I raised. In the lack of such an understanding, the race toward developing state-of-the-art systems risks becoming a race toward tuning systems to the test set. It invalidates the notion of the existence of a true and independent gold standard. This is actually where we started: annotated corpora—in which the understanding of the phenomenon is reflected—determine the subsequent sequence of events.

1.5.3 New directions

Having observed the drawbacks above, I felt that in order for the coreference task to be feasible and, most importantly, useful for the NLP community, several foundational aspects had to be reconsidered. This was a challenging endeavor and far beyond the scope of this work, yet I was unavoidably induced to take the first (three) opening steps in this direction, which are presented in this section. The first two steps occurred along the way as a direct result of ongoing research efforts; the third one was more ambitious but necessary to complete this thesis.

CISTELL The entity-mention system built by Luo et al. (2004) opened a new, promising way to solve coreference, yet it remains “an area that needs further research,” as they point out. The approach taken here adds to the body of work on entity-mention models by devising a system, CISTELL, that handles discourse entities as (growing) baskets (Chapter 4).[9] The notion of a growing basket is akin to Heim’s (1983) *file card* in file change semantics, where a file card stands for each

[9] *Cistell* is the Catalan word for ‘basket.’

Mention: ***his professional colleagues***	
Attribute	**Value**
Is a pronoun	false
Head	colleagues
Is-a	co-worker, fellow worker
Hypernym	associate
Gender	—
Number	plural
Specifier	his
Counter	—
NE type	—
Modifiers	professional
Sentence head	be
Grammatical role	oblique
Word position	17
Mention position	6
Sentence position	2

Table 1.4: Information contained in a basket

discourse entity so that the information of subsequent references can be stored in it as the discourse progresses.

At the beginning of the discourse, a basket is allocated to each mention, containing the mention attributes such as head, type, modifiers, etc., as exemplified in Table 1.4. Some information is directly extracted from the text, and some from external resources like WordNet. The convenient property of baskets is that they can *grow* by swallowing other baskets and incorporating their attributes. A sketch of this process is displayed in Fig. 1.1, which shows a document in the middle of being processed with baskets symbolically represented. When two baskets are classified as coreferent, they are immediately clustered into a growing basket (which can grow further).

The general resolution process is inspired by Popescu-Belis et al. (1998). Fig. 1.2 is a still illustration of the growing process, showing the different moves that CISTELL can make at a given point. The highlighted basket is the one under consideration. It can either be swallowed and contribute to making the big basket on the left grow further (Fig. 1.2(a)); or be swallowed by one of the smaller growing baskets on the right (Fig. 1.2(b)); or stay a singleton, with a chance of growing later in the discourse (Fig. 1.2(c)).

The crux of the matter is the growing process, namely to decide whether or not two baskets should be merged for growing. The decision takes on a different form depending on whether the two baskets are singletons or whether one of them has already started to grow. The former decision is straightforward as it is only based on the coreference probability given by the pairwise classifier, while the latter can be made in several ways taking into consideration the different pairwise decisions

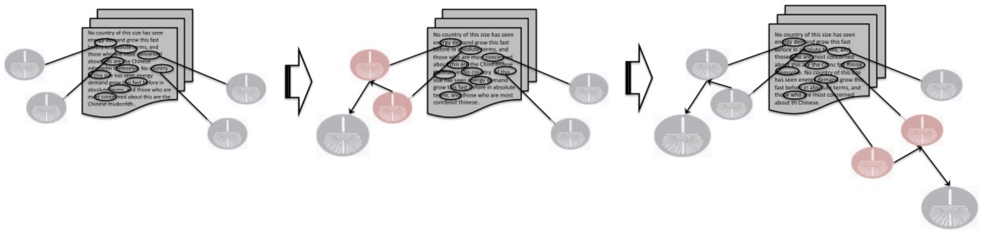


Figure 1.1: Depiction of CISTELL's coreference resolution process

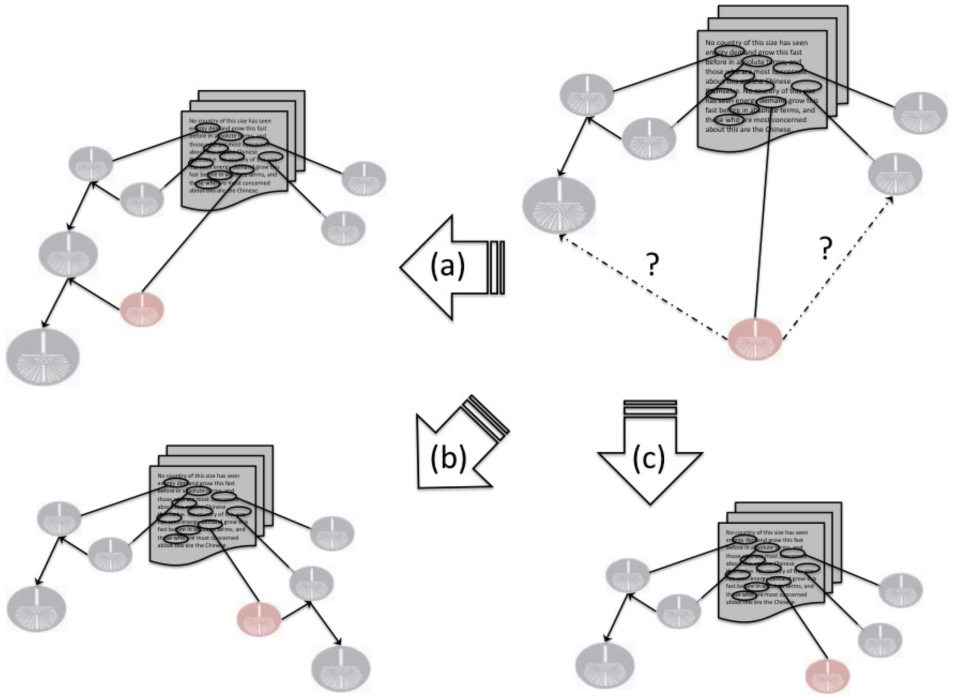


Figure 1.2: Depiction of CISTELL's basket growing process. (a) Growing to the left, (b) Growing to the right, (c) Singleton.

with each of the baskets in the growing basket. I call each such decision a *match*. A parameter is defined that specifies the number of matches required for a basket to be swallowed by a growing entity. Of the different values that were tried (any-one-match, any-two-matches, . . . , all-matches), the best results were obtained when all the matches had to be coreferentially positive. This requirement is referred to as *strong match*. On the opposite end, allowing any positive match to be a guarantee for the merging to occur (i.e., *weak match*) results in overmerging.

Contrary to Luo et al. (2004) then, the better performance of the strong-match (versus weak-match) strategy provides evidence of the beneficial effects of using a more global strategy. Keeping track of the history of each discourse entity is helpful to capture the largest amount of information about an entity provided by the text. On the other hand, however, global strategies can have adverse effects if not properly designed. Distance features, for instance, only make sense if the notion of antecedence is taken into account (Kehler, 1997). Thus, they work for pronouns and full NPs, but can introduce noise in the case of proper nouns, which are more sensitive to other features such as entity frequency (Haghighi and Klein, 2010).

In principle, baskets are unlimited and can be enriched with as many attributes as appropriate and available—with information from "inside" the text as well as background and world knowledge from "outside" the text—thus making it possible to encode many of the features needed to solve coreference relations. The coreference classifier is jointly trained for coreference resolution and discourse-new detection. This is achieved by generating negative training instances that, unlike Soon et al. (2001), include not only coreferent mentions but also singletons. Although it was not feasible to exploit CISTELL to its full potential given the obstacles that were encountered along the way (Section 1.5.2), it provides a framework of possibilities to accommodate the theoretical turn that I introduce below (see "Coreference continuum").

BLANC To overcome the observed shortcomings of the widely-used measures MUC, B^3, and CEAF, I drew on the Rand index (Rand, 1971) to devise BLANC (BiLateral Assessment of Noun-phrase Coreference), a new measure that takes not only coreference but also non-coreference links into account and, most importantly, balances them equally (Chapter 5). In this way, singletons are neither ignored nor given greater importance than multi-mention entities regardless of their frequency of occurrence. 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.

In BLANC's default setting, the two types of links (coreference and non-coreference) count the same for the final score, which enhances the weight of multi-mention entities. Yet BLANC is defined as a weighted measure with a parameter α that allows the user to decide whether more weight should be given to coreference or non-coreference links. Splitting the reward in half between coref-

	MUC			B^3			CEAF	BLANC		
	P	R	F	P	R	F	P / R / F	P	R	blanc
AnCora - Spanish										
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 - English										
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

Table 1.5: BLANC results for Table 4.3, Chapter 4.
1. = ALL SINGLETONS; 2. = HEAD MATCH; 3. = HEAD MATCH + PRON; 4. = STRONG MATCH; 5. = SUPER STRONG MATCH; 6. = BEST MATCH; 7. = WEAK MATCH

erence and non-coreference links means that an upper threshold of 50% recall is set to each type of link. This directly penalizes systems that output a large number of singletons and too few coreference links, or, in the extreme case, "coreference" systems that do nothing but return singletons. Not ignoring any of the two link types, and achieving the most satisfactory balance between them, are two main desiderata behind the definition of BLANC. The measure was tested on different corpora ranging from ACE—with its restricted semantic types—to OntoNotes and AnCora.

Since BLANC is a new measure, I did not include it in the experiments with CISTELL reported in Chapter 4, yet it is worthwhile to do it here. Tables 1.5, 1.6, and 1.7 reproduce Tables 4.3, 4.5, and 4.7, but include the BLANC scores next to MUC, B^3, and CEAF. It emerges that BLANC clearly stands apart from B^3 and CEAF in relation to the all-singletons baseline (row number 1). Consequently, the rest of BLANC scores are always below B^3 and usually below or slightly above CEAF due to the lack of the initial boost from singleton identification. On the other end, and unlike MUC, BLANC also punishes overmerging quite notably (WEAK MATCH, row number 7, tends to include many—correct and incorrect—links).

Interestingly enough, B^3 and CEAF never agree in the best ranked system, while BLANC and CEAF do two out of six times. It appears from the scatterplots in Fig. 1.3 that corpus should be distinguished as a relevant variable to find significant correlations between the measures, especially in the case of BLANC. There is a positive correlation between BLANC and CEAF in both ACE ($\tau = 0.82$)[10] and

[10] All correlations are measured using Kendall's tau (τ) and are significant at $p < 0.01$.

	MUC			B^3			CEAF	BLANC		
	P	R	F	P	R	F	P / R / F	P	R	blanc
OntoNotes scheme										
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
ACE scheme										
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

Table 1.6: BLANC results for Table 4.5, Chapter 4.
1. = All singletons; 2. = Head match; 3. = Head match + pron; 4. = Strong match; 5. = Super strong match; 6. = Best match; 7. = Weak match

	MUC			B^3			CEAF	BLANC		
	P	R	F	P	R	F	P / R / F	P	R	blanc
OntoNotes scheme										
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
ACE scheme										
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

Table 1.7: BLANC results for Table 4.7, Chapter 4.
1. = All singletons; 2. = Head match; 3. = Head match + pron; 4. = Strong match; 5. = Super strong match; 6. = Best match; 7. = Weak match

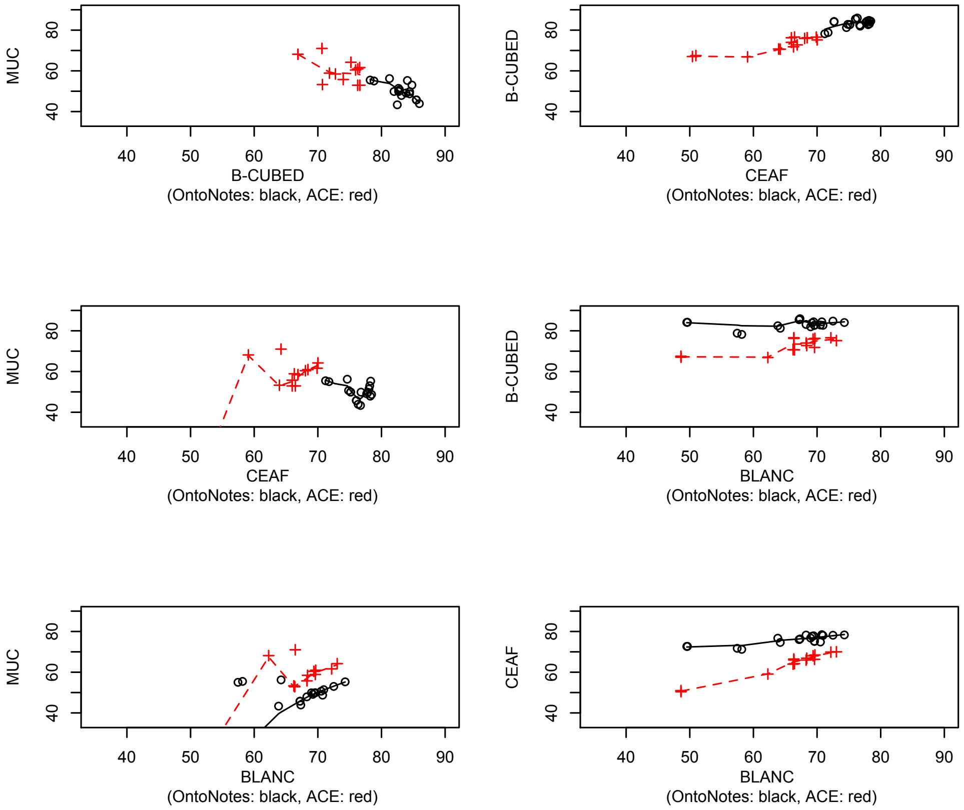


Figure 1.3: Pairwise scatterplots of MUC, B^3, CEAF and BLANC

OntoNotes ($\tau = 0.57$), between BLANC and MUC in both ACE ($\tau = 0.53$) and OntoNotes ($\tau = 0.48$), and to a lesser degree between BLANC and B^3 but only in ACE ($\tau = 0.43$). CEAF is also positively correlated in ACE with B^3 ($\tau = 0.62$) and MUC ($\tau = 0.46$). The correlation between BLANC and the other measures is lower in OntoNotes as this corpus has a larger number of singletons. As I discussed, the distinguishing characteristic of BLANC is its treatment of singletons. It is therefore more desirable than the other measures to compare coreference performances on corpora that are annotated with the complete set of mentions.

The quantitative analysis is supported by a qualitative evaluation. Appendix A provides a sample of system outputs for the same two documents from OntoNotes. For the first document (nbc_0030, Appendix A.1), all four measures rank the Strong match output as the best, but disagree in the ranking of the other three CISTELL's outputs. B^3 and CEAF, as opposed to MUC, clearly dislike Weak match due to their bias toward singletons. BLANC scores the non-Strong-match outputs very similarly and at a much lower range, thus hinting at their inferior quality.

For the second document (voa_0207, Appendix A.2), the winning output, for being the less noisy, is SUPER STRONG MATCH and it is also the one scored highest by all four measures. Again, MUC stands alone in ranking WEAK MATCH high. As evidence for the higher discriminative power of BLANC, STRONG MATCH and BEST MATCH are given the same score by CEAF, while BLANC shows a slight preference for the former that agrees with a qualitative analysis of the two outputs.

The BLANC measure was publicly used for the first time in the SemEval competition, as supported by the goals of the task (Chapter 6). Apart from the outputs of CISTELL, Appendix A.1 also includes the outputs of the six participating systems. It is then possible to compare the numerical scores presented in Table 6.5 with the automatically annotated texts. From a qualitative perspective, the best output is that by CORRY-C, which is also the one best ranked according to CEAF and BLANC. In contrast, B^3 ranks RELAXCOR first, whose output includes a very small number of coreference links, and MUC rewards SUCRE, which tends to produce too many links. This provides further evidence for the singleton bias of B^3, and the overmerging bias of MUC, respectively. TANL-1 and UBIU are ranked last by most metrics and this accords with the poor quality of their outputs. Finally, it is not possible to compare BART with the other outputs on the same level because it only participated in the regular setting and so scores for true mentions are not available.

Coreference continuum It is surprising that so few studies (Versley, 2008; Poesio and Artstein, 2005; Charolles and Schnedecker, 1993) have considered the theoretical implications and problems with the current definition of coreference. If even humans do not agree on what is and what is not a coreference relation, unrealistic expectations on the performance of automatic coreference resolution systems should not be imposed. As already pointed out in Section 1.5.2, there are a number of relations in real data—as exemplified by (12) and (13)—that cannot be accounted for in terms of either coreference or non-coreference.

I identify the source of the problem as the assumption that coreference is explicable in terms of strict, categorical identity. Instead, I introduce the notion of *near-identity* in conjunction with a continuum model of coreference (Chapter 8). This provides room for understanding and modeling cases in which the identity relation is not total but partial: X and Y are the same with regard to particular features, but differ with regard to at least one feature. In everyday language, we often utter statements like *You and I are the same* when the intended meaning is that you and I have many things in common, or occupy a similar social position, although we are obviously two different persons and are different in many respects.

Moving into the domain of coreference, it is suggested that the (non-)coreference judgments we make are directly connected to the level of granularity in which the entities of a particular discourse are categorized. This level of granularity is in turn determined by the communicative purposes and coherence of the discourse.

As an example, consider the categorization of *Spain* in two different contexts: A historical description will favor a non-coreferential reading of *Spain* and *the modern Spain*, while a different context, like a current news article, will probably use the two indistinctly, favoring a coreferential reading.

Positing a coreference continuum is consistent with the various degrees at which referentiality seems to operate (Fraurud, 1992). In fact, the more referential an entity is, the more we can specify it. While it makes perfect sense to conceive of Spain at a general, synchronous level, or to split it into *17th-century Spain*, *modern Spain*, etc., it would be odd to think of the fork that you are eating with in terms of *yesterday's fork* or *tomorrow's fork*. The fact that not all entities can be conceived at the same level of specificity brings us back to my earlier point about the role of ontological types in entity individuation. Just as categories are organized along a continuum, coreference occurs along a continuum from identity to non-identity, with a range of near-identity relations between these extremes.

In line with Bybee (2010), who considers that linguistic structure derives from the application of domain-general cognitive processes (e.g., categorization, chunking), I suggest three cognitive operations of categorization that account for near-identity relations holding between entities that share most, but not all, feature values. Depending on whether a discourse entity augments, overrides or nullifies a feature value of an existing entity, I distinguish between *specification*, *refocusing* and *neutralization*, respectively. The first two create new indexical features by showing a different facet of a complex entity; whereas the latter conflates two or more similar entities into a single category, thus neutralizing potential indexical features. This three-way distinction provides a flexible framework that, among other benefits, offers an operational rationale for the treatment of metonymy from the perspective of coreference, which remains a controversial area and major cause of inter-annotator disagreement.

Specification, refocusing and neutralization are represented within the framework of Fauconnier's (1985; 1997) theory of mental spaces. It is a conceptually appropriate framework and it also fits well with the ideas behind the CISTELL system. Mental spaces are abstract mental structures that we construct while we think and talk for purposes of local understanding, and onto which discourse entities are projected. Fauconnier recognizes that the tools of formal logic fail when confronted with the full range of natural language phenomena. Mental spaces, but not formal logic, can explain cases such as the split self (19) and split coreference (20) whose meaning requires splitting a single referent into two.

(19) a. If [I] were you, I'd hate [me].
b. If [I] were you, I'd hate [myself].

(20) If [Woody Allen] had been born twins, [they] would have been sorry for each other, but [he] wasn't and so he's only sorry for himself.

Entities in a discourse are conceptualized by discourse participants with a set of associated features that have specific values according to the particular space. When a new entity is introduced with properties that either clash with or remove detail from an existing DE, then a new mental space with the corresponding near-identical entity needs to be built. Split coreference is one of the relations that require a mental space shift. I have identified ten feature types that require such shifts when their values are changed, and organized them into a hierarchy (Recasens et al., 2010a). This hierarchy provides the most frequent ways that give rise to near-identity relations. The details come toward the end of the thesis in Chapter 8. Appendix B supplements that chapter by providing the collection of near-identity excerpts that were used in the inter-annotator agreement study leading to the near-identity typology, and Appendix C reports the annotators' classification of the highlighted relations in the excerpts.

1.6 Major contributions

The primary contribution of this thesis is the critical but also constructive insight on various aspects of the coreference resolution task, ranging from resolution and evaluation to corpus annotation and theory. The final outcome is the development of a new approach to coreference based on the synthesis of corpus data, domain-general cognitive operations and Fauconnier's (1985; 1997) mental space theory. Unlike the prevailing either-or approach, the continuum model of coreference that I propose allows for middle-ground relations of near-identity to account satisfactorily for naturally occurring data.

The major contributions of this thesis are:

- The Spanish and Catalan AnCora corpora (totaling nearly 800k words) annotated with coreference information, together with a coding scheme that addresses shortcomings of previous approaches and that incorporates specific tags for Spanish and Catalan.[11]

- A list of over forty-five learning features that are tested in a pairwise model of coreference resolution in Spanish. They are shown to be weakly informative on their own but support complex and unpredictable interactions.

- An entity-mention model of coreference resolution called CISTELL that combines and integrates pairwise decisions at the discourse level.

- An exposition of weaknesses in the widely-used coreference evaluation measures that obscure the evaluation of coreference systems and hinder direct comparisons between state-of-the-art systems.

- The implementation of the Rand index for coreference evaluation in the new BLANC measure, which takes into account—equally by default—both coreference and non-coreference links.

[11] http://clic.ub.edu/corpus/en/ancora

- A comparison between coreference and paraphrase, highlighting both their similarities and their differences in order to suggest areas of mutual collaboration between coreference resolution and paraphrase extraction.
- The organization of a SemEval shared task on coreference resolution, and the public release on the Web of the datasets, scorers, and documentation that were developed for the task.[12]
- A small corpus of 60 near-identity extracts, a typology of near-identity relations and an inter-annotator agreement study that proves its stability.
- A continuum model of coreference, ranging from identity to non-identity through near-identity relations, and three cognitive operations of categorization (i.e., specification, neutralization and refocusing) that account for the different stages along this continuum.
- The identification of a number of weaknesses of the current approach to coreference resolution that suggest a need to reconsider various aspects of the task.

These contributions constitute the contents of the following Chapters 2 to 8.

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

Part I

CORPUS ANNOTATION WITH COREFERENCE

CHAPTER 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 and 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 and 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., 2009a). Due to the complexity inherent in coreference, limitations of rule-based approaches (Hobbs, 1978; Baldwin, 1997; Lappin and 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 and Cardie, 2002b; 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 and 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 and 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, 2004a; Hovy et al., 2006; Poesio and Artstein, 2008).

Coreferentially annotated corpora are even scarcer for languages other than English. Among these few we find Czech, German and Dutch (Kučová and 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 and 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 and 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., 2007b). 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 and Chinchor, 1997; Doddington et al., 2004); the MATE meta-scheme (Davies et al., 1998; Poesio, 2004b) 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, 2004b; Poesio and 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

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

[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.

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.'

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).

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

(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.

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.

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

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 Fuerzas Armadas de este país "consiguieron destruir buena parte de las fuerzas convencionales de UNITA*." El general sudafricano hizo estas declaraciones.

[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'	*∅*	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)

> '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 and 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., 2007b). 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.

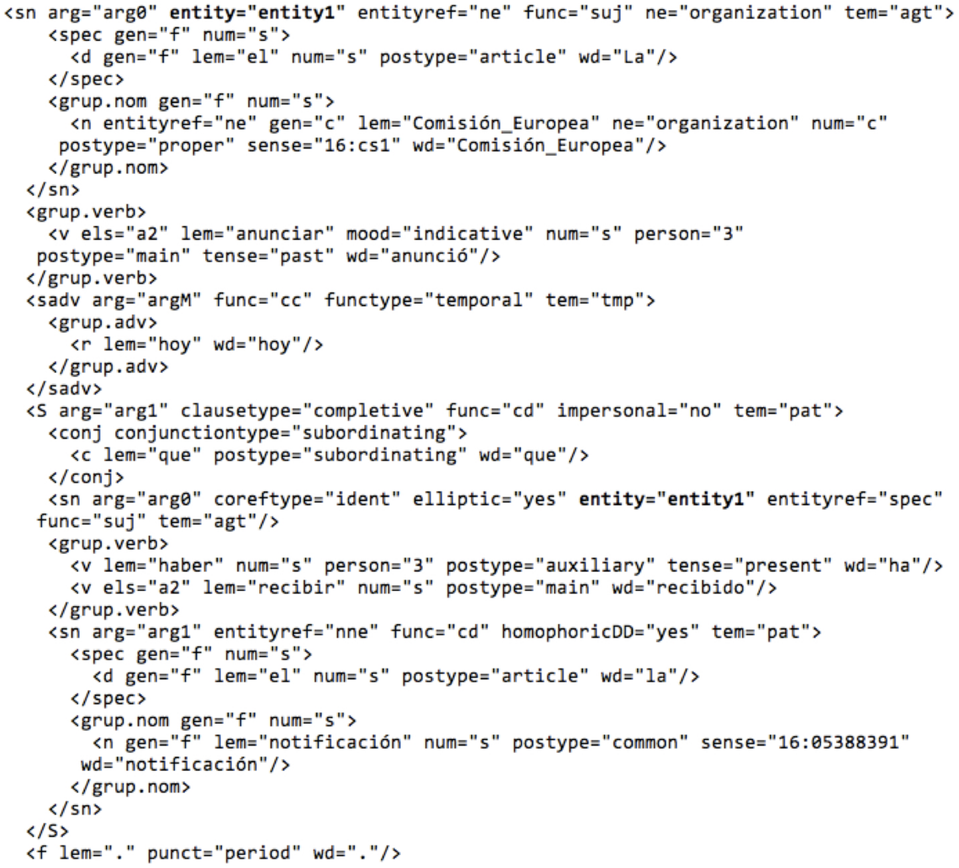
(17) a. (Sp.) Unión Fenosa Inversiones es *una empresa del grupo español Unión Fenosa*.
'Unión Fenosa Inversiones is *a company in the Spanish group Unión Fenosa*.'
b. (Cat.) Han demanat una entrevista amb el conseller d'Indústria, *Antoni Subirà*.
'They have asked for an interview with the Minister 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]
- 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

[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="."/>
```

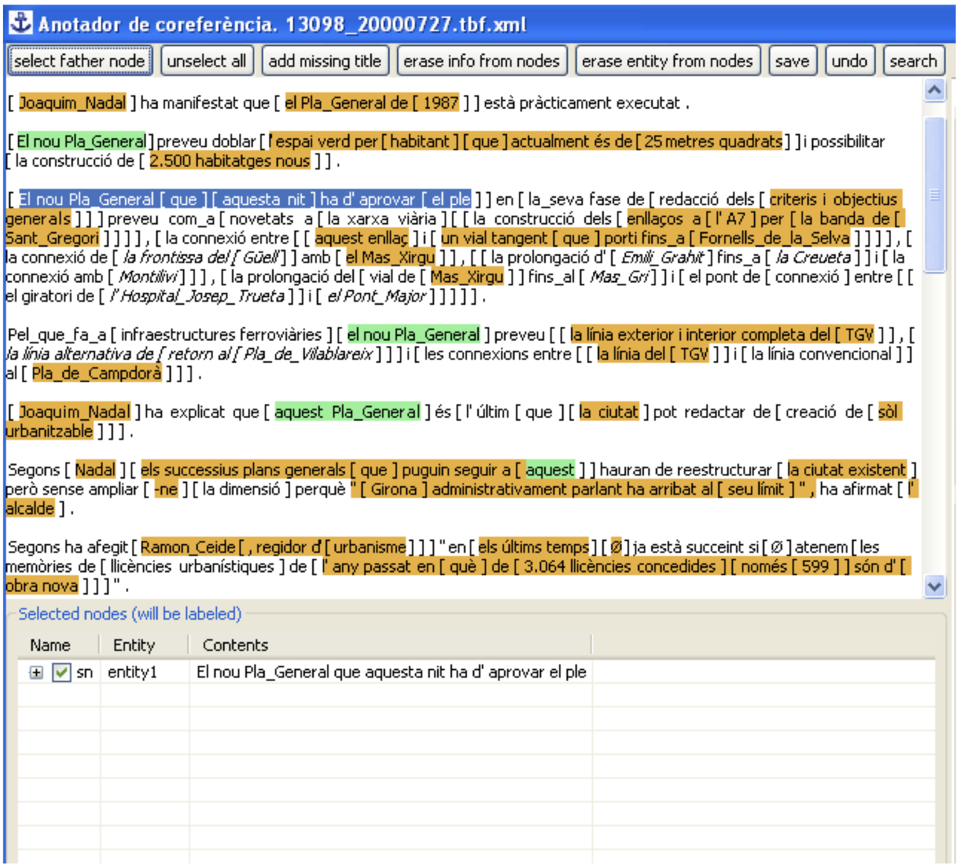
Figure 2.1: The XML file format exemplified with the sentence *La Comisión Europea anunció hoy que ϕ ha recibido la notificación* ‘The European Commission announced today that (it) received the notification.’ Notice that the bold entity number “entity1” marks the identity coreference relation between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

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*.’

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



[Joaquim_Nadal] ha manifestat que [el Pla_General de [1987]] està pràcticament executat .

[El nou Pla_General] preveu doblar [l' espai verd per [habitant] [que] actualment és de [25 metres quadrats]] i possibilitar [la construcció de [2.500 habitatges nous]] .

[El nou Pla_General [que] [aquesta nit] ha d' aprovar [el ple]] en [la_seva fase de [redacció dels [criteris i objectius generals]]] preveu com_a [novetats a [la xarxa viària] [[la construcció dels [enllaços a [l' A7] per [la banda de [Sant_Gregori]]]] , [la connexió entre [[aquest enllaç] i [un vial tangent [que] porti fins_a [Fornells_de_la_Selva]]]] , [la connexió de [*la frontissa del* [*Güell*]] amb [el Mas_Xirgu]] , [[la prolongació d' [*Emili_Grahit*] fins_a [*la Creueta*]] i [la connexió amb [*Montilivi*]]] , [la prolongació del [vial de [Mas_Xirgu]] fins_al [*Mas_Gri*]] i [el pont de [connexió] entre [[el giratori de [*l' Hospital_Josep_Trueta*]] i [*el Pont_Major*]]]]] .

Pel_que_fa_a [infraestructures ferroviàries] [el nou Pla_General] preveu [[la línia exterior i interior completa del [TGV]] , [*la línia alternativa de* [*retorn al* [*Pla_de_Vilablareix*]]] i [les connexions entre [[la línia del [TGV]] i [la línia convencional]] al [Pla_de_Campdorà]]] .

[Joaquim_Nadal] ha explicat que [aquest Pla_General] és [l' últim [que] [la ciutat] pot redactar de [creació de [sòl urbanitzable]]] .

Segons [Nadal] [els successius plans generals [que] puguin seguir a [aquest]] hauran de reestructurar [la ciutat existent] però sense ampliar [-ne] [la dimensió] perquè " [Girona] administrativament parlant ha arribat al [seu límit] " , ha afirmat [l' alcalde] .

Segons ha afegit [Ramon_Ceide [, regidor d' [urbanisme]]] " en [els últims temps] [Ø] ja està succeint si [Ø] atenem [les memòries de [llicències urbanístiques] de [l' any passat en [què] de [3.064 llicències concedides] [només [599]] són d' [obra nova]]] " .

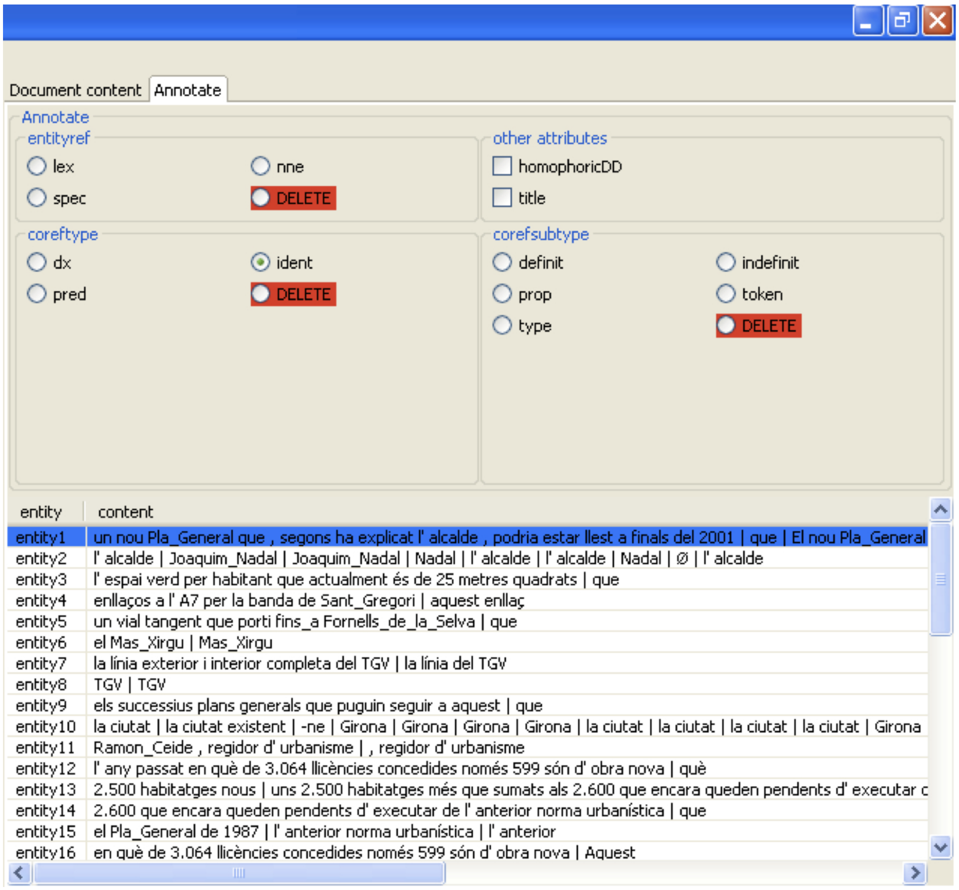
Name	Entity	Contents
⊞ ☑ sn	entity1	El nou Pla_General que aquesta nit ha d' aprovar el ple

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

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 and 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 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.



Document content | Annotate

Annotate

entityref: ○ lex ○ nne ○ spec ○ DELETE

other attributes: □ homophoricDD □ title

coreftype: ○ dx ● ident ○ pred ○ DELETE

corefsubtype: ○ definit ○ indefinit ○ prop ○ token ○ type ○ DELETE

<table>
<tr><th>entity</th><th>content</th></tr>
<tr><td>entity1</td><td>un nou Pla_General que , segons ha explicat l' alcalde , podria estar llest a finals del 2001 | que | El nou Pla_General</td></tr>
<tr><td>entity2</td><td>l' alcalde | Joaquim_Nadal | Joaquim_Nadal | Nadal | l' alcalde | l' alcalde | Nadal | Ø | l' alcalde</td></tr>
<tr><td>entity3</td><td>l' espai verd per habitant que actualment és de 25 metres quadrats | que</td></tr>
<tr><td>entity4</td><td>enllaços a l' A7 per la banda de Sant_Gregori | aquest enllaç</td></tr>
<tr><td>entity5</td><td>un vial tangent que porti fins_a Fornells_de_la_Selva | que</td></tr>
<tr><td>entity6</td><td>el Mas_Xirgu | Mas_Xirgu</td></tr>
<tr><td>entity7</td><td>la línia exterior i interior completa del TGV | la línia del TGV</td></tr>
<tr><td>entity8</td><td>TGV | TGV</td></tr>
<tr><td>entity9</td><td>els successius plans generals que puguin seguir a aquest | que</td></tr>
<tr><td>entity10</td><td>la ciutat | la ciutat existent | -ne | Girona | Girona | Girona | Girona | la ciutat | la ciutat | la ciutat | la ciutat | Girona</td></tr>
<tr><td>entity11</td><td>Ramon_Ceide , regidor d' urbanisme | , regidor d' urbanisme</td></tr>
<tr><td>entity12</td><td>l' any passat en què de 3.064 llicències concedides només 599 són d' obra nova | què</td></tr>
<tr><td>entity13</td><td>2.500 habitatges nous | uns 2.500 habitatges més que sumats als 2.600 que encara queden pendents d' executar c</td></tr>
<tr><td>entity14</td><td>2.600 que encara queden pendents d' executar de l' anterior norma urbanística | que</td></tr>
<tr><td>entity15</td><td>el Pla_General de 1987 | l' anterior norma urbanística | l' anterior</td></tr>
<tr><td>entity16</td><td>en què de 3.064 llicències concedides només 599 són d' obra nova | Aquest</td></tr>
</table>

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

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 and Vieira (1998) in Swedish and English, respectively. The number of first-mention definites in Spanish and Catalan is even higher (see Recasens et al. (2009b) 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 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.

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 (%)

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 and Poesio (2008) point out the main problems in using percent agreement and the kappa coefficient (Siegel and 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, 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 of two texts. In the last two sessions, they annotated the two test texts separately from each other.

[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 and 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)

2.7.1.5 Results

Artstein and 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 and 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 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

	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)

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 and Poesio (2008), it is “the lack of consensus

[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)

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 (1980) suggests α=.8 as a threshold value, which is supported by more recent efforts (Artstein and 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

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 α = .87 ([.857,.898] 95% CI) for Text 1, and α = .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 interpretations. 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 attributive 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 relations 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

COREFERENCE RESOLUTION AND EVALUATION

CHAPTER 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 and 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 and Cardie, 2002b) seem to be a promising way to combine and weigh the relevant factors, overcoming the limitations of constraint-based approaches (Lappin and 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 and 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 and Cardie (2002b) 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 and Cardie (2002b): backward features describing the antecedent of the candidate antecedent (Yang et al., 2004), semantic information from Wikipedia, WordNet and semantic roles (Ponzetto and 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 linguistic

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

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 overrepresented in her feature set.

Bengtson and 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 and 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 and 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 and Cardie, 2002b; 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 and 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 and 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 and 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 and 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 and 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 and 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.

CHAPTER 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., 2007a), and AnCora (Recasens and 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 and 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 and Roth, 2008; Denis and Baldridge, 2009), or on both (Finkel and Manning, 2008; Haghighi and 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 and 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., 2007a; 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 and 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 and Manning, 2008; Poon and Domingos, 2008; Haghighi and Klein, 2009; Ng, 2009). It is not our aim to compete with them. Rather, we conduct three

[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., 2010b).

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 and 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 with TiMBL (Daelemans and Bosch, 2005). Basket-growing is done in four differ-

[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.’

ent 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 and Cardie (2002b)'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]

4.3.3 Features

We follow Soon et al. (2001), Ng and Cardie (2002b) and Luo et al. (2004) to generate most of the 29 features we use for the pairwise model. These include

[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.

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

[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

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.

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 and Baldridge, 2009). They are consistent across the two corpora though: MUC rewards WEAK MATCH the most, B^3

	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

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.

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

[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

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.

(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

[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

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 and 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 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).

		#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

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 and 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, 1998). Syntactic chunks were obtained from YamCha 1.33, an SVM-based NP-chunker (Kudoh and 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

	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

the same, the latter contains fewer words as it lacks the null elements (traces, ellipsed 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.

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.

[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.

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., 2010b). 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).

CHAPTER 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 and 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 and 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 and Marcu, 2005); and a third group includes two or more scores for the sake of completeness (Luo et al., 2004; Bengtson and Roth, 2008; Ng, 2009; Finkel and Manning, 2008; Poon and 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.

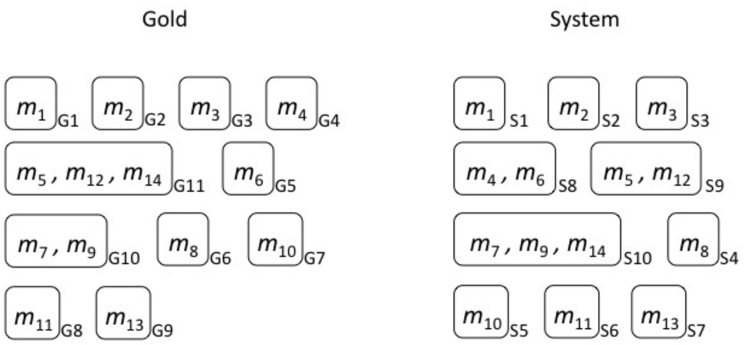


Figure 5.2: The problem of comparing the gold partition with the system partition for a given text (Fig. 5.1)

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

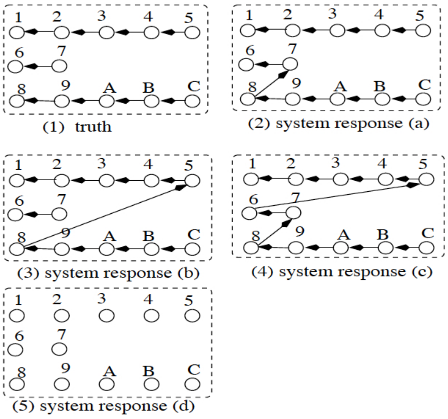


Figure 5.3: Example entity partitions (from Luo (2005))

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 and 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 and 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 and 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 and 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 and 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 and Cardie, 2007; Poon and Domingos, 2008; Haghighi and Klein, 2009). Pairwise F1 simply computes P, R, and F over all pairs of coreferent mentions. As noted by Haghighi and 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 and SYS correspond to the entropy of GOLD and SYS, respectively. Then the

[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.

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 and 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 and 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 and Zitouni (2005) as the best systems for each data set. This sort of discrepancy makes

[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 and Manning (2008)	64.1	73.8		
Poon and Domingos (2008)	68.4	69.2	63.9	
Denis and 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 and Zitouni (2005)	86.0	83.7	82.0	91.6
Haghighi and Klein (2007)	63.3			
Bengtson and Roth (2008)	75.8	80.8	75.0	
Poon and Domingos (2008)	69.1	71.2	65.9	
Wick and McCallum (2009)	70.1	81.5		

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

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 and 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 to have clean clusters plus a cluster with miscellaneous items over having clusters

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.

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 cor-

rectly 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 large the class is; assigning a mention to a small class is scored equally as assigning

[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.

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	$\mathrm{P}_c = \frac{rc}{rc+wc}$	$\mathrm{P}_n = \frac{rn}{rn+wn}$	$\text{BLANC-P} = \frac{\mathrm{P}_c+\mathrm{P}_n}{2}$
R	$\mathrm{R}_c = \frac{rc}{rc+wn}$	$\mathrm{R}_n = \frac{rn}{rn+wc}$	$\text{BLANC-R} = \frac{\mathrm{R}_c+\mathrm{R}_n}{2}$
F	$\mathrm{F}_c = \frac{2\mathrm{P}_c\mathrm{R}_c}{\mathrm{P}_c+\mathrm{R}_c}$	$\mathrm{F}_n = \frac{2\mathrm{P}_n\mathrm{R}_n}{\mathrm{P}_n+\mathrm{R}_n}$	$\text{BLANC} = \frac{\mathrm{F}_c+\mathrm{F}_n}{2}$

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 and 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:

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

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 $\mathrm{BLANC}_{\alpha=0.1}$ 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 and Roth (2008) simply discard twinless mentions, and Rahman and Ng (2009) limit to removing only those twinless system mentions that are singletons, as in these cases no penalty should be applied. Recently, Cai and 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 and 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	BLANC$_{\alpha=0.5}$	BLANC$_{\alpha=0.2}$	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 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 and Cardie (2002b), 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 and 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 and Zitouni (2005)	86.0	83.7	82.0	91.6	**81.4**
Bengtson and 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., 2010b), 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.

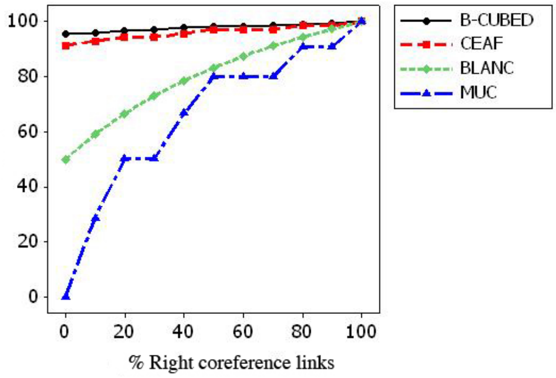


Figure 5.6: The BLANC score curve as the number of right coreference links increases

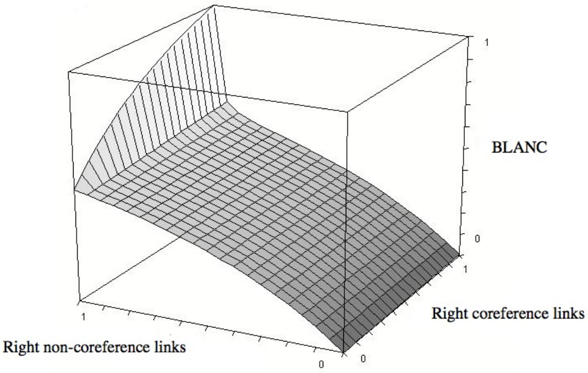


Figure 5.7: The BLANC score surface as a function of right coreference and right non-coreference links, for data from Table 5.8

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).

CHAPTER 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$^\diamond$, and Yannick Versley‡

*University of Barcelona
** Technical University of Catalonia
† University College Ghent
$^\diamond$ 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 and 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 and 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 and 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 and 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., 2007a) 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 and Laws, 2008), and dependency information by MaltParser (Hall and 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 and 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 and 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 and 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 and 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 differences between corpora, there are other factors that might explain this rank-

[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

ing: (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

COREFERENCE THEORY

CHAPTER 7

On Paraphrase and Coreference

Marta Recasens and Marta Vila

University of Barcelona

To appear as a squib in *Computational Linguistics*, 36(4):639–647

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 and 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	
		✓	✗
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. (2010a) 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 and Chinchor, 1997; Doddington et al., 2004), hence the criticism by van Deemter and 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 relations 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 and 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 and 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).

CHAPTER 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 and 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., 2007a) 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. Entering the domain of language, we will consider the language-specific counterparts

[1] Appendix B includes the kind of real data that traditional models have failed to explain and that we build on for this article.

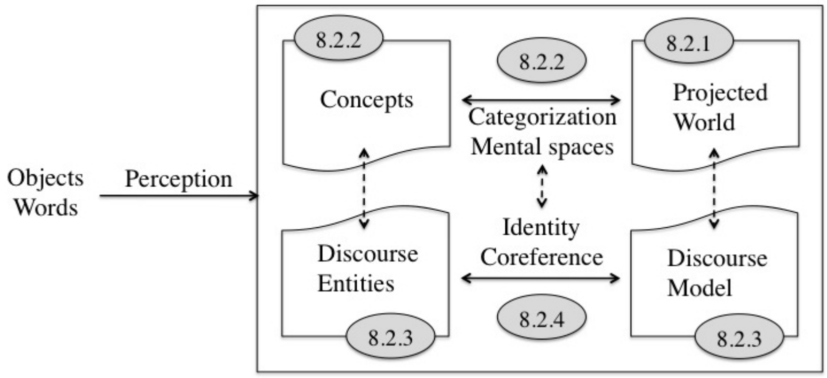


Figure 8.1: Language and cognition

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 unprincipled variability. The fact that the processes by which we construct the

[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).

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 importer is actually the very ugly Lord Lipton, then (3-a) is true, whereas (3-b) is

[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 $x = y$.

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 and 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" (Nunberg, 1984:207).

[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.

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.

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

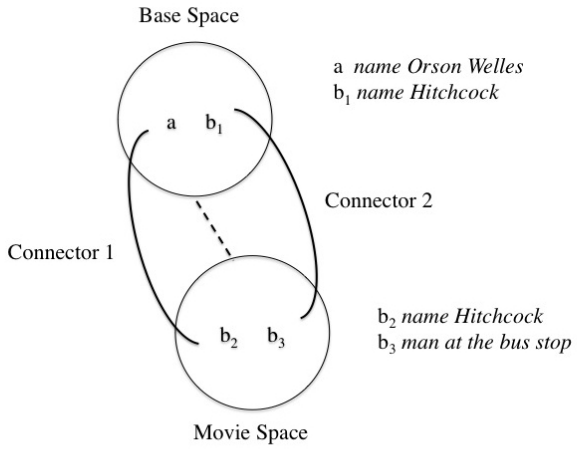


Figure 8.2: Mental space configuration of (4)

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 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

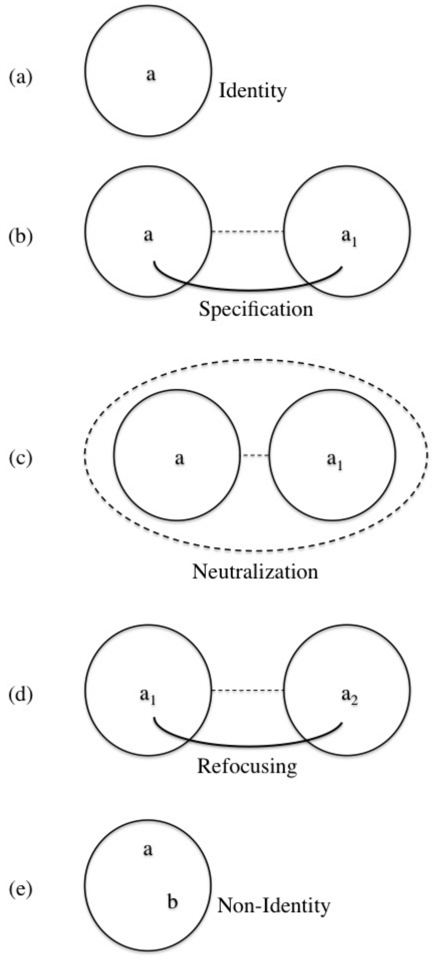


Figure 8.3: Mental space configurations of the coreference continuum

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.

(6) Last night in Tel Aviv, *Jews* attacked a restaurant that employs Palestinians. "*We* want war," *the crowd* chanted.

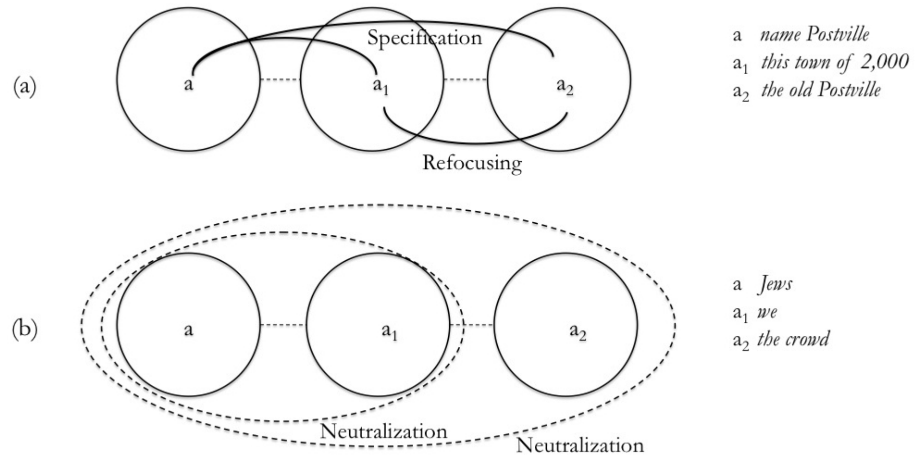


Figure 8.4: Mental space configurations of (5) and (6)

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., 2010a). 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 Fig. 8.5(b), where a and a_1 are only related by specification and no third mental space needs to be introduced.

[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.

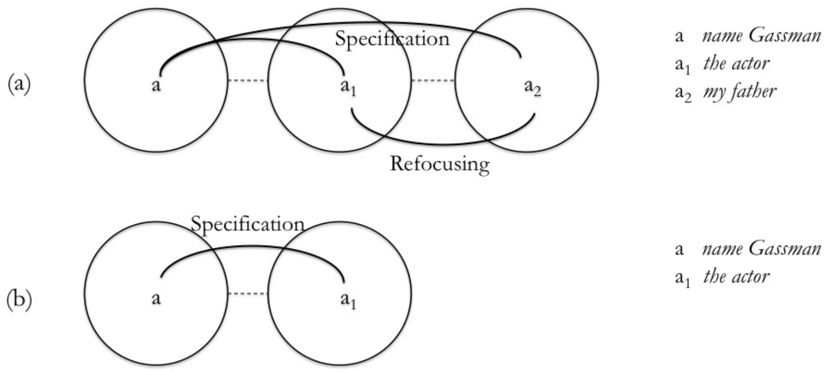


Figure 8.5: Mental space configurations of (7-a) and (7-b)

(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).

(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.

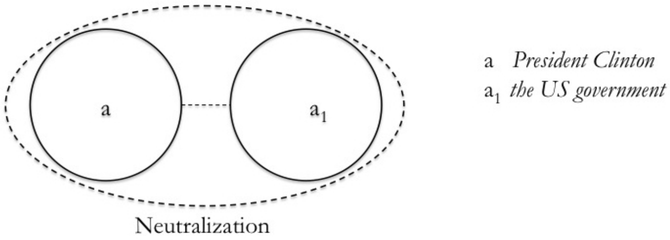


Figure 8.6: Mental space configuration of (12)

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., 2007a) and AnCora (Recasens and 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 annotation guidelines was used, since the typology underwent substantial revision—in a decreasing manner—after completing each round.

[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.

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 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.

[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.

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.

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.

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

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).

⋆ ⋆ ⋆

CHAPTER 9

Conclusions and Future Directions

In this final chapter, I look back at what has been accomplished in this thesis. The main accomplishment has been to expand the understanding of coreference by taking a broader view of the problem and uncovering various barriers that currently hinder the successful performance of coreference resolution systems. This thesis has advanced the understanding of noteworthy issues associated with the theoretical approach, corpus annotation, computational treatment, and evaluation of the coreference problem. I have recast the problem in slightly different terms to bring it closer to the linguistic reality.

The chapter begins by briefly highlighting the main contributions of this work and assessing the lessons I have gained (Section 9.1), and then presents some interesting and challenging ideas emerged from my analysis that need to be tackled in future research (Section 9.2).

9.1 Conclusions

This thesis offers a number of contributions beyond previous work to different facets of the coreference problem. Broken down into facets, I have:

- Annotation
    - Developed coreference annotation guidelines for Catalan and Spanish data.
    - Built AnCora-CO, the largest Catalan and Spanish corpora annotated with coreference relations.[1]

[1] http://clic.ub.edu/corpus/en/ancora

- Resolution
  - Gathered over forty-five learning features for coreference resolution in Spanish and analyzed their contribution in a pairwise model.
  - Presented CISTELL, a state-of-the-art coreference resolution system that allows for using discourse and background knowledge as well as cluster-level features.
  - Organized and developed the resources for the first SemEval-2010 shared task on "Coreference Resolution in Multiple Languages."[2]
- Evaluation
  - Carried out a broad comparative analysis between different evaluation metrics for coreference: MUC, B^3, CEAF, ACE-Value, Pairwise F1, Mutual information, and the Rand index.
  - Developed the BLANC measure for coreference evaluation that provides a solution to the problem of singletons exhibited by the other measures. It adapts the Rand index to reward coreference and non-coreference links equally.
- Theory
  - Showed the similarities and differences between the concepts of coreference and paraphrase.
  - Argued for the need to introduce the concept of *near-identity* in the usual discrete analysis of coreference.
  - Built the first corpus of real texts containing cases of near-identity, defined a typology of near-identity relations, and conducted an inter-annotator agreement study.
  - Presented a novel model of coreference based on a continuum and three cognitive operations of categorization: specification, neutralization and refocusing.

The contributions of this thesis are valuable in that they shed light on assumptions underlying the vast majority of past research and clarify the issues to be resolved. As far as annotation is concerned, it has been argued that it needs to be backed up by a **more comprehensive theory of coreference**. We need a theory that explains the complex patterns encountered in naturally occurring data such as relations that do not fall neatly into either coreference or non-coreference, and mentions that are on the bridge between referentiality and non-referentiality. The continuum model of coreference that I have presented offers an appropriate theoretical framework for considering the non-discrete nature of coreference and, to some extent, of referentiality as well. At present, the lack of a **true gold-standard corpus** has serious

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

implications for the development of coreference resolution systems and, most importantly, for their comparison. I have advocated for annotation efforts that identify the set of mentions with the entire set of referential NPs—thus excluding attributive and predicative phrases—and that annotate multi-mention entities as well as singletons.

A second limitation of existing approaches is associated with the shortcomings of the currently available learning features, which turn out to be hardly generalizable and to lack **pragmatic and background knowledge**. Encoding this knowledge in some form that a coreference resolution system can use is key to substantially advancing the state of the art. Given the wide range of environments in which coreference relations can occur, taking into account the context rather than the mentions in isolation is also essential. The CISTELL system that I have developed gives the opportunity to store and carry along the resolution process not only the information about a mention provided "inside" the text, but also background and world knowledge from "outside" the text. Only in this way can we succeed in ensuring that the system is given enough information to decide, for each individual mention, whether it does or does not refer to a specific entity. Although learning-based methods can be of great help during this process, they should be **manually engineered** rather than blindly applied.

Finally, the major problems with the current coreference evaluation practices result from **biases in the scoring metrics widely in use today** (B^3, CEAF, MUC). As a consequence of this, for instance, the all-singletons and head-match baselines are difficult to beat when systems are tested on a corpus annotated with the entire set of mentions. This is aggravated by the fact that state-of-the-art systems do not tend to be evaluated qualitatively nor provide their outputs. The measure that I have proposed, BLANC, aims to find a **trade-off between the large number of singletons and the relatively small number of multi-mention entities**. In addition, it is argued that evaluating mention identification and coreference resolution based on a single score can be misleading in some situations. Instead, I have suggested using true mentions and separating **mention identification as a task in itself**. Last but not least, scores alone do not suffice to measure the performance of a system, and making **system outputs** publicly available, as it has been done with the SemEval participating systems, should become common practice.

Overall, it is hoped that the insights provided in this thesis will ultimately help find more successful ways to approach the coreference resolution task. I have already made some steps in this direction and have ideas for further work that I would like to pursue in the future. This is the focus of the next section.

9.2 Future directions

As a result of the discussions held with several researchers and of the analysis carried out over the course of this project, a number of directions in which to extend the work presented here have emerged. This section outlines the most significant ones, some of which have already been mentioned throughout the body of the dissertation. Two questions stand out as the largest challenges facing coreference today: (1) the world of discourse entities, their representation and behavior; and (2) the nature and operation of a dynamic and non-discrete coreference system.

Regarding the world of discourse entities, the continuum model of coreference has many advantages including coverage and robustness. Further work is needed in determining which features and dimensions play an important role, as governed by the semantics and the context, as well as in determining the feature values from the text, the context and background knowledge (e.g., the Web, Wikipedia, databases, etc.). This is the information that should be captured in the baskets used by the CISTELL system in order to include the kind of background and world knowledge that coreference systems are lacking today. As Ng (2010) points out in his survey, unsupervised approaches rival their supervised counterparts and this casts doubts on "whether supervised resolvers are making effective use of the available labeled data." The problem arises from the fact that texts do not make explicit all the information that is required for their understanding (but that people recover effortlessly). A reliable way to get this information is by drawing on recent work on extracting knowledge from the Web (Markert and Nissim, 2005; Kozareva and Hovy, 2010) and on machine reading (Peñas and Hovy, 2010). Although extracting and integrating all information from a single text is still beyond current capabilities, it should be thought of as a long-term goal.

From a linguistic point of view, there is much room for debate as to what exactly a *context* is. The notion is usually parameterized according to empirical or theoretical aims. For example, Bach (2005) explains that "what is loosely called 'context' is the conversational setting broadly construed. It is the mutual cognitive context, or salient common ground. It includes the current state of the conversation (what has just been said, what has just been referred to, etc.), the physical setting (if the conversants are face to face), salient mutual knowledge between the conversants, and relevant broader common knowledge." Therefore, another interesting long-term program of research is toward a comprehensive theory of context.

Further insight into the continuum model of coreference and the near-identity idea can be gained by carrying out psycholinguistic experiments: Does near-identity have an effect in processing time? To what extent does the context determine one reading or another? Do the operations of specification, neutralization and refocusing have any psycholinguistic reality? As a support to as well as an extension of psycholinguistic research, developing a large corpus annotated with near-identity relations is a short-term goal that will make it possible to study specific issues such as the interaction of near-identity with the temporal structure and directionality of discourse, or the interaction of near-identity with the entity's ontological type. In addition, it will provide training and test data for developing more refined coreference resolution systems, and encourage other researchers to work in this same direction. The typology presented in Chapter 8 can serve as a starting point to formulate the annotation guidelines.

Fully integrating the near-identity continuum and the CISTELL system leads to the other large question of building a dynamic and non-discrete coreference resolution system. A rich vein of research in this realm lies in establishing the principles which should allow a coreference engine to make its decisions. This requires fleshing out the mental space diagrams of Chapter 8 in more objective and measurable terms, and formalizing the operations of neutralization, refocusing, etc. Typed feature structures and unification may prove helpful for this purpose.

After fifteen years of learning-based coreference research, it has become clear that the mention-pair model is weak and that global models offer better performance (Ng, 2010), but it is still unclear what is the best approach to design cluster-level features and combine their information. Developing a system that handles baskets not statically but dynamically is a logical starting point. Eventually we should arrive at a system that is able to use contextual and world knowledge as well as the interlocutors' intentions to automatically infer which features are potentially important and so which feature values to propagate to new mental spaces and which to not propagate. To address this, baskets should be represented not as a fixed feature set with values in or out, but as one with probabilistic feature value membership that supports much more nuanced matching.

The way and order in which text, context and world knowledge should be exploited in building baskets is closely related to the way in which the contents encoded in different baskets should be compared. Ideally then, work on the choice and value definition of basket features and work on basket matching should be collaborative and, to the extent possible, coordinated.

Finally, regarding evaluation issues, the use of the BLANC measure in future studies to report coreference scores will, over time, result in further refinements like the proper adjustment of the alpha parameter. I have shown BLANC's main strengths and weaknesses with respect to the current measures, but a number of issues associated with the definition of the formulas of all measures used in coreference resolution, such as considering whether corrections for chance are needed (Vinh et al., 2009), or examining typical variances of scores under different conditions and data sizes, await further study. Research on this field will be greatly enhanced if the coreference community adopts a standard evaluation metric in the immediate future.

After having provided both answers and questions, the work presented in this thesis ends here. I have answered a number of questions, but also raised new ones that should stimulate further research in the fascinating area of coreference.

Bibliography

Abad, A., Bentivogli, L., Dagan, I., Giampiccolo, D., Mirkin, S., Pianta, E., and Stern, A. (2010). A resource for investigating the impact of anaphora and coreference on inference. In *Proceedings of LREC 2010*, pages 128–135, Valletta, Malta.

Alshawi, H. (1990). Resolving quasi logical forms. *Computational Linguistics*, 16(3):133–144.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Aone, C. and Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of ACL 1995*, pages 122–129.

Appelt, D., Hobbs, J., Bear, J., Israel, D., Kameyama, M., and Tyson, M. (1995). SRI International FASTUS System MUC-6 Results and Analysis. In *Proceedings of MUC-6*, Columbia, Maryland.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24(1):65–87.

Ariel, M. (2001). Accessibility Theory: An overview. In Sanders, T., Schliperoord, J., and Spooren, W., editors, *Text Representation*, pages 29–87. John Benjamins, Amsterdam.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76:B13–B26.

Arnold, J. E. and Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56:521–536.

Artstein, R. and Poesio, M. (2005). Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL 2005*, pages 141–150, Edinburgh, UK.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Attardi, G., Rossi, S. D., and Simi, M. (2010). TANL-1: Coreference resolution by parse analysis and similarity clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 108–111, Uppsala, Sweden.

Azzam, S., Humphreys, K., and Gaizauskas, R. (1999). Using coreference chains for text summarization. In *Proceedings of the ACL Workshop on Coreference and its Applications*, pages 77–84, Baltimore, Maryland.

Bach, K. (2005). Context ex machina. In Szabó, Z. G., editor, *Semantics versus Pragmatics*, pages 15–44. Clarendon, Oxford.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistic Coreference*, pages 563–566, Granada, Spain.

Baker, M. C. (2003). *Lexical Categories*. Cambridge University Press, Cambridge.

Baldwin, B. (1997). CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL-EACL Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts*, pages 38–45, Madrid.

Barbu, C., Evans, R., and Mitkov, R. (2002). A corpus based analysis of morphological disagreement in anaphora resolution. In *Proceedings of LREC 2002*, pages 1995–1999, Las Palmas de Gran Canaria, Spain.

Barker, C. (2010). Nominals don't provide criteria of identity. In Rathert, M. and Alexiadou, A., editors, *The Semantics of Nominalizations across Languages and Frameworks*, pages 9–24. Mouton de Gruyter, Berlin.

Bean, D. L. and Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of ACL 1999*, pages 373–380, College Park, Maryland.

Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proceedings of EMNLP 2008*, pages 294–303, Honolulu, Hawaii.

Berger, A., Pietra, S. D., and Pietra, V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bergsma, S., Lin, D., and Goebel, R. (2008). Distributional identification of non-referential pronouns. In *Proceedings of ACL-HLT 2008*, pages 10–18, Columbus, Ohio.

Bertran, M., Borrega, O., Recasens, M., and Soriano, B. (2008). AnCoraPipe: A tool for multilevel annotation. *Procesamiento del Lenguaje Natural*, 41:291–292.

Bhagat, R. (2009). *Learning Paraphrases from Text*. PhD thesis, University of Southern California, Los Angeles, California.

Blackwell, S. (2003). *Implicatures in Discourse: The Case of Spanish NP Anaphora*. John Benjamins, Amsterdam.

Borrega, O., Taulé, M., and Martí, M. A. (2007). What do we mean when we talk about named entities? In *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham, UK.

Bosque, I. and Demonte, V., editors (1999). *Gramática descriptiva de la lengua española*. Real Academia Española / Espasa Calpe, Madrid.

Boyd, A., Gegg-Harrison, W., and Byron, D. (2005). Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated features. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, pages 40–47, Ann Arbor, Michigan.

Brants, T. (2000). TnT – A statistical part-of-speech tagger. In *Proceedings of ANLP 2000*, Seattle, Washington.

Broscheit, S., Poesio, M., Ponzetto, S. P., Rodríguez, K. J., Romano, L., Uryupina, O., Versley, Y., and Zanoli, R. (2010). BART: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 104–107, Uppsala, Sweden.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press, New York.

Byron, D. K. (2001). The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics*, 27(4):569–578.

Byron, D. K. and Gegg-Harrison, W. (2004). Eliminating non-referring noun phrases from coreference resolution. In *Proceedings of DAARC 2004*, pages 21–26, Azores, Portugal.

Cai, J. and Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of SIGDIAL*, pages 28–36, University of Tokyo, Japan.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, DOI:10.1007/s10579-010-9120-1.

Callison-Burch, C. (2007). *Paraphrasing and Translation.* PhD thesis, University of Edinburgh, Edinburgh, UK.

Carbonell, J. and Brown, R. G. (1988). Anaphora resolution: a multi-strategy approach. In *Proceedings of COLING 1988*, pages 96–101, Budapest.

Cardie, C. and Wagstaff, K. (1999). Noun phrase coreference as clustering. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, College Park, Maryland.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carreiras, M. and Gernsbacher, M. A. (1992). Comprehending conceptual anaphors in Spanish. *Language and Cognitive Processes*, 7(3&4):281–299.

Chaffin, R., Herrmann, D. J., and Winston, M. (1988). An empirical taxonomy of part-whole relations: Effects of part-whole relation type on relation identification. *Language and Cognitive Processes*, 3(1):17–48.

Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Proceedings of ACL 2008*, pages 789–797, Columbus, Ohio.

Charolles, M. and Schnedecker, C. (1993). Coréférence et identité: le problème des référents évolutifs. *Langages*, 112:106–126.

Choi, Y. and Cardie, C. (2007). Structured local training and biased potential functions for conditional random fields with application to coreference resolution. In *Proceedings of HLT-NAACL 2007*, pages 65–72, Rochester, New York.

Clark, H. H. (1977). Bridging. In Johnson-Laird, P. and P.C.Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.

Connolly, D., Burger, J. D., and Day, D. S. (1994). A machine learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP-1)*, pages 255–261, Manchester, UK.

Crawley, R., Stevenson, R., and Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 4:245–264.

Culotta, A., Wick, M., Hall, R., and McCallum, A. (2007). First-order probabilistic models for coreference resolution. In *Proceedings of HLT-NAACL 2007*, pages 81–88, Rochester, New York.

Daelemans, W. and Bosch, A. V. (2005). *Memory-Based Language Processing*. Cambridge University Press, Cambridge.

Daelemans, W., Buchholz, S., and Veenstra, J. (1999). Memory-based shallow parsing. In *Proceedings of CoNLL 1999*, Bergen, Norway.

Daumé III, H. and Marcu, D. (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of HLT-EMNLP 2005*, pages 97–104, Vancouver, Canada.

Davies, S., Poesio, M., Bruneseaux, F., and Romary, L. (1998). Annotating coreference in dialogues: Proposal for a scheme for MATE. `http://www.ims.uni-stuttgart.de/projekte/mate/mdag`.

Denis, P. (2007). *New learning models for robust reference resolution*. PhD thesis, University of Texas at Austin, Austin, Texas.

Denis, P. and Baldridge, J. (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of NAACL-HLT 2007*, Rochester, New York.

Denis, P. and Baldridge, J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of EMNLP 2008*, pages 660–669, Honolulu, Hawaii.

Denis, P. and Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation. In *Proceedings of LREC 2004*, pages 837–840, Lisbon.

Dolan, B., Brockett, C., and Quirk, C. (2005). README file included in the Microsoft Research Paraphrase Corpus. Redmond, Washington.

Dras, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD thesis, Macquarie University, Sydney, Australia.

Eckert, M. and Strube, M. (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Elsner, M. and Charniak, E. (2010). The same-head heuristic for coreference. In *Proceedings of ACL 2010 Short Papers*, pages 33–37, Uppsala, Sweden.

Erk, K. and Strapparava, C., editors (2010). *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden.

Evans, R. (2000). A comparison of rule-based and machine learning methods for identifying non-nominal *it*. In *Proceedings of NLP 2000*, volume 1835/2000 of *LNAI*, pages 233–241, Berlin. Springer-Verlag.

Fauconnier, G. (1985). *Mental Spaces: Aspects of Meaning Construction in Natural Language*. MIT Press, Cambridge.

Fauconnier, G. (1997). *Mappings in Thought and Language*. Cambridge University Press, Cambridge.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

Ferrández, A., Palomar, M., and Moreno, L. (1999). An empirical approach to Spanish anaphora resolution. *Machine Translation*, 14:191–216.

Finkel, J. R. and Manning, C. D. (2008). Enforcing transitivity in coreference resolution. In *Proceedings of ACL-HLT 2008*, pages 45–48, Columbus, Ohio.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York, 2nd edition.

Frampton, M., Fernández, R., Ehlen, P., Christoudias, M., Darrell, T., and Peters, S. (2009). Who is "you"? Combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of EACL 2009*, pages 273–281, Athens.

Fraurud, K. (1990). Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, 7:395–433.

Fraurud, K. (1992). *Processing Noun Phrases in Natural Discourse*. PhD thesis, Stockholm University, Stockholm.

Fraurud, K. (1996). Cognitive ontology and NP form. In Fretheim, T. and Gundel, J. K., editors, *Reference and Referent Accessibility*, pages 65–87. John Benjamins, Amsterdam.

Frege, G. (1892). On sense and reference. In Geach, P. and Black, M., editors, *Translations from the Philosophical Writings of Gottlob Frege*, pages 56–78. Basil Blackwell (1952), Oxford.

Fuchs, C. (1994). *Paraphrase et énonciation. Modélisation de la paraphrase langagière*. Ophrys, Paris.

Fujita, A. (2005). *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. PhD thesis, Nara Institute of Science and Technology, Ikoma, Nara, Japan.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., and Wilks, Y. (1995). Description of the LaSIE system as used for MUC-6. In *Proceedings of MUC-6*, pages 207–220.

Gamer, M., Lemon, J., and Fellows, I. (2009). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.82.

Garigliano, R., Urbanowicz, A., and Nettleton, D. J. (1997). University of Durham: Description of the LOLITA system as used in MUC-7. In *Proceedings of MUC-7*.

Geach, P. (1962). *Reference and Generality*. Cornell University Press, Ithaca.

Geach, P. (1967). Identity. *Review of Metaphysics*, 21:3–12.

Gerber, M. and Chai, J. Y. (2010). Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of ACL 2010*, pages 1583–1592.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: a brief history. In *Proceedings of COLING 1996*, pages 466–471, Copenhagen.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

GuoDong, Z. and Fang, K. (2009). Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *Proceedings of EMNLP 2009*, pages 978–986, Suntec, Singapore.

Haghighi, A. and Klein, D. (2007). Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of ACL 2007*, pages 848–855, Prague.

Haghighi, A. and Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP 2009*, pages 1152–1161, Suntec, Singapore.

Haghighi, A. and Klein, D. (2010). Coreference resolution in a modular, entity-centered model. In *Proceedings of HLT-NAACL 2010*, pages 385–393, Los Angeles, California.

Hall, J., Nilsson, J., Nivre, J., Eryigit, G., Megyesi, B., Nilsson, M., and Saers, M. (2007). Single malt or blended? A study in multilingual parser optimization. In *Proceedings of the shared task session of CoNLL 2007*, pages 933–939, Prague.

Hall, J. and Nivre, J. (2008). A dependency-driven parser for German dependency and constituency representations. In *Proceedings of the ACL Workshop on Parsing German (PaGe 2008)*, pages 47–54, Columbus, Ohio.

Halliday, M. A. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hammami, S., Belguith, L., and Hamadou, A. B. (2009). Arabic anaphora resolution: Corpora annotation with coreferential links. *The International Arab Journal of Information Technology*, 6(5):481–489.

Harabagiu, S., Bunescu, R., and Maiorano, S. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of NAACL 2001*, pages 55–62.

Hasler, L., Orasan, C., and Naumann, K. (2006). NPs for events: Experiments in coreference annotation. In *Proceedings of LREC 2006*, pages 1167–1172, Genoa, Italy.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Heim, I. (1983). File change semantics and the familiarity theory of definiteness. In Bäuerle, R., Schwarze, C., and von Stechow, A., editors, *Meaning, Use, and Interpretation of Language*, pages 164–189. Mouton de Gruyter, Berlin.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Vloet, J. V. D., and Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Hervás, R. and Finlayson, M. (2010). The prevalence of descriptive referring expressions in news and narrative. In *Proceedings of ACL 2010 Short Papers*, pages 49–54, Uppsala, Sweden.

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and Trushkina, J. (2004). Recent developments in Linguistic Annotations of the TüBa-D/Z Treebank. In *Proceedings of TLT 2004*, Tübingen, Germany.

Hinrichs, E. W., Filippova, K., and Wunsch, H. (2007). A data-driven approach to pronominal anaphora resolution in German. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, pages 115–124. John Benjamins, Amsterdam.

Hinrichs, E. W., Kübler, S., and Naumann, K. (2005). A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20, Ann Arbor, Michigan.

Hirschman, L. and Chinchor, N. (1997). MUC-7 Coreference Task Definition – Version 3.0. In *Proceedings of MUC-7*.

Hirst, G. J. (1981). *Anaphora in natural language understanding: a survey*. Springer-Verlag, Berlin.

Hobbs, J. (1985). Granularity. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI 1985)*, pages 432–435, Los Angeles, California.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44:311–338.

Hoste, V. (2005). *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, University of Antwerp, Antwerp, Belgium.

Hoste, V. and De Pauw, G. (2006). KNACK-2002: A richly annotated corpus of Dutch written text. In *Proceedings of LREC 2006*, pages 1432–1437, Genoa, Italy.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of HLT-NAACL 2006*, pages 57–60, New York.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Ide, N. (2000). Searching annotated language resources in XML: A statement of the problem. In *Proceedings of the ACM SIGIR Workshop on XML and Information Retrieval*, Athens.

Iida, R., Inui, K., Takamura, H., and Matsumoto, Y. (2003). Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pages 23–30, Budapest.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the ACL Workshop on Linguistic Annotation*, pages 132–139, Prague.

Jackendoff, R. (1983). *Semantics and Cognition*. MIT Press, Cambridge.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

Kabadjov, M. A. (2007). *A comprehensive evaluation of anaphora resolution and discourse-new classification*. PhD thesis, University of Essex, Colchester, UK.

Kameyama, M. (1998). Intrasentential centering: A case study. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering Theory in Discourse*, pages 89–112. Oxford University Press, Oxford.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J., Janssen, T., and Stokhof, M., editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematical Centre Tracts 135, Amsterdam.

Karttunen, L. (1976). Discourse referents. In McCawley, J., editor, *Syntax and Semantics*, volume 7, pages 363–385. Academic Press, New York.

Kehler, A. (1997). Probabilistic coreference in information extraction. In *Proceedings of EMNLP 1997*, pages 163–173, Providence, Rhode Island.

Kehler, A., Appelt, D., Taylor, L., and Simma, A. (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of NAACL 2004*, pages 289–296, Boston, Massachusetts.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25(1):1–44.

Kennedy, C. and Boguraev, B. (1996). Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of COLING 1996*, pages 113–118, Copenhagen.

Kilgarriff, A. (1999). 95% Replicability for manual word sense tagging. In *Proceedings of EACL 1999*, pages 277–278, Bergen, Norway.

Klenner, M. and Ailloud, É. (2009). Optimization in coreference resolution is not needed: A nearly-optimal algorithm with intensional constraints. In *Proceedings of EACL 2009*, pages 442–450, Athens.

Kobdani, H. and Schütze, H. (2010). SUCRE: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 92–95, Uppsala, Sweden.

Kozareva, Z. and Hovy, E. (2010). Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of ACL 2010*, pages 1482–1491, Uppsala, Sweden.

Krahmer, E. (2010). What computational linguists can learn from psychologists (and vice versa). *Computational Linguistics*, 36(2):285–294.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest Studies in Philosophy*, 2:255–276.

Krippendorff, K. (2004 [1980]). *Content Analysis: An Introduction to its Methodology*. Sage, Thousand Oaks, California, second edition. Chapter 11.

Kudoh, T. and Matsumoto, Y. (2000). Use of support vector learning for chunk identification. In *Proceedings of CoNLL 2000 and LLL 2000*, pages 142–144, Lisbon.

Kučová, L. and Hajičová, E. (2004). Coreferential relations in the Prague Dependency Treebank. In *Proceedings of DAARC 2004*, pages 97–102, Azores, Portugal.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Lasersohn, P. (2000). *Same*, models and representation. In Jackson, B. and Mathews, T., editors, *Proceedings of Semantics and Linguistic Theory 10*, pages 83–97, Cornell, New York. CLC Publications.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP 2005*, pages 25–32, Vancouver, Canada.

Luo, X. (2007). Coreference or not: A twin model for coreference resolution. In *Proceedings of HLT-NAACL 2007*, pages 73–80, Rochester, New York.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of ACL 2004*, pages 21–26, Barcelona.

Luo, X. and Zitouni, I. (2005). Multi-lingual coreference resolution with syntactic features. In *Proceedings of HLT-EMNLP 2005*, pages 660–667, Vancouver, Canada.

Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V. B., and Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006*, pages 963–968, Genoa, Italy.

Markert, K. and Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–401.

Mayol, L. and Clark, R. (2010). Pronouns in Catalan: Games of partial information and the use of linguistic resources. *Journal of Pragmatics*, 42:781–799.

McCallum, A. and Wellner, B. (2005). Conditional models of identity uncertainty with application to noun coreference. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 905–912, Cambridge, Massachusetts. MIT Press.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *Proceedings of IJCAI 1995*, pages 1050–1055, Montreal, Canada.

Milićević, J. (2007). *La paraphrase*. Peter Lang, Bern.

Mirkin, S., Berant, J., Dagan, I., and Shnarch, E. (2010). Recognising entailment within discourse. In *Proceedings of COLING 2010*, Beijing.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of COLING-ACL 1998*, pages 869–875, Montreal, Canada.

Mitkov, R. (2002). *Anaphora Resolution*. Longman, London.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., and Sotirova, V. (2000). Coreference and anaphora: Developing annotating tools, annotated resources and annotation strategies. In *Proceedings of DAARC 2000*, pages 49–58, Lancaster, UK.

Mitkov, R. and Hallett, C. (2007). Comparing pronoun resolution algorithms. *Computational Intelligence*, 23(2):262–97.

Morton, T. S. (1999). Using coreference in question answering. In *Proceedings of TREC-8*, pages 85–89, Gaithersburg, Maryland.

Morton, T. S. (2000). Coreference for NLP applications. In *Proceedings of ACL 2000*, pages 173–180, Hong Kong.

Müller, C. (2007). Resolving *it*, *this* and *that* in unrestricted multi-party dialog. In *Proceedings of ACL 2007*, pages 816–823, Prague.

Müller, C., Rapp, S., and Strube, M. (2002). Applying co-training to reference resolution. In *Proceedings of ACL 2002*, pages 352–359, Philadelphia, Pennsylvania.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt.

Navarretta, C. (2004). Resolving individual and abstract anaphora in texts and dialogues. In *Proceedings of COLING 2004*, pages 233–239, Geneva, Switzerland.

Navarretta, C. (2007). A contrastive analysis of abstract anaphora in Danish, English and Italian. In *Proceedings of DAARC 2007*, pages 103–109, Lagos, Portugal.

Navarretta, C. (2009a). Automatic recognition of the function of singular neuter pronouns in texts and spoken data. In Devi, S. L., Branco, A., and Mitkov, R., editors, *Anaphora Processing and Applications (DAARC 2009)*, volume 5847 of *LNAI*, pages 15–28, Berlin / Heidelberg. Springer-Verlag.

Navarretta, C. (2009b). Co-referential chains and discourse topic shifts in parallel and comparable corpora. *Procesamiento del Lenguaje Natural*, 42:105–112.

Navarro, B. (2007). *Metodología, construcción y explotación de corpus anotados semántica y anafóricamente*. PhD thesis, University of Alicante, Alicante, Spain.

Ng, V. (2003). Machine learning for coreference resolution: Recent successes and future challenges. Technical report CUL.CIS/TR2003-1918, Cornell University, New York.

Ng, V. (2004). Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of ACL 2004*, pages 151–158, Barcelona.

Ng, V. (2005). Machine learning for coreference resolution: from local classification to global ranking. In *Proceedings of ACL 2005*, pages 157–164, Ann Arbor, Michigan.

Ng, V. (2007). Shallow semantics for coreference resolution. In *Proceedings of IJCAI 2007*, pages 1689–1694, Hyderabad, India.

Ng, V. (2008). Unsupervised models for coreference resolution. In *Proceedings of EMNLP 2008*, pages 640–649, Honolulu, Hawaii.

Ng, V. (2009). Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of NAACL-HLT 2009*, pages 575–583, Boulder, Colorado.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of ACL 2010*, pages 1396–1411, Uppsala, Sweden.

Ng, V. and Cardie, C. (2002a). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of COLING 2002*, pages 1–7, Taipei.

Ng, V. and Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, pages 104–111, Philadelphia, Pennsylvania.

Nicolae, C. and Nicolae, G. (2006). BestCut: a graph algorithm for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 275–283, Sydney, Australia.

Nicolae, C., Nicolae, G., and Roberts, K. (2010). C-3: Coherence and coreference corpus. In *Proceedings of LREC 2010*, pages 136–143, Valletta, Malta.

Nicolov, N., Salvetti, F., and Ivanova, S. (2008). Sentiment analysis: Does coreference matter? In *Proceedings of the Symposium on Affective Language in Human and Machine*, Aberdeen, UK.

Nilsson, K. (2010). *Hybrid Methods for Coreference Resolution in Swedish*. PhD thesis, Stockholm University, Stockholm.

Nunberg, G. (1984). Individuation in context. In *Proceedings of the 2nd West Coast Conference on Formal Linguistics (WCCFL 2)*, pages 203–217, Stanford, California.

Orasan, C. (2003). PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 39–43, Sapporo, Japan.

Orasan, C., Cristea, D., Mitkov, R., and Branco, A. (2008). Anaphora Resolution Exercise: An overview. In *Proceedings of LREC 2008*, pages 2801–2805, Marrakech, Morocco.

Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., and Muñoz, R. (2001). An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4):545–567.

Passonneau, R. (2004). Computing reliability for coreference annotation. In *Proceedings of LREC 2004*, pages 1503–1506, Lisbon.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of LREC 2006*, pages 831–836, Genoa, Italy.

Peñas, A. and Hovy, E. (2010). Semantic enrichment of text with background knowledge. In *Proceedings of the NAACL First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 15–23, Los Angeles, California.

Poesio, M. (2004a). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona.

Poesio, M. (2004b). The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop at HLT-NAACL 2004*, pages 154–162, Boston, Massachusetts.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Poesio, M., Delmonte, R., Bristot, A., Chiran, L., and Tonelli, S. (2004a). The VENEX corpus of anaphora and deixis in spoken and written Italian. Manuscript. Available online at `http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf`.

Poesio, M., Kruschwitz, U., and Chamberlain, J. (2008). ANAWIKI: Creating anaphorically annotated resources through Web cooperation. In *Proceedings of LREC 2008*, pages 2352–2355, Marrakech, Morocco.

Poesio, M., Ponzetto, S. P., and Versley, Y. (forthcoming). Computational models of anaphora resolution: A survey. *Linguistic Issues in Language Technology*.

Poesio, M., Stevenson, R., Eugenio, B. D., and Hitzeman, J. (2004b). Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Poesio, M., Uryupina, O., and Versley, Y. (2010). Creating a coreference resolution system for Italian. In *Proceedings of LREC 2010*, pages 713–716, Valletta, Malta.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of HLT-NAACL 2006*, pages 192–199, New York.

Poon, H., Christensen, J., Domingos, P., Etzioni, O., Hoffmann, R., Kiddon, C., Lin, T., Ling, X., Mausam, Ritter, A., Schoenmackers, S., Soderland, S., Weld, D., Wu, F., and Zhang, C. (2010). Machine Reading at the University of Washington. In *Proceedings of the NAACL-HLT First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 87–95, Los Angeles, California.

Poon, H. and Domingos, P. (2008). Joint unsupervised coreference resolution with Markov logic. In *Proceedings of EMNLP 2008*, pages 650–659, Honolulu, Hawaii.

Popescu-Belis, A. (2000). Évaluation numérique de la résolution de la référence: critiques et propositions. *T.A.L.: Traitement automatique de la langue*, 40(2):117–146.

Popescu-Belis, A., Rigouste, L., Salmon-Alt, S., and Romary, L. (2004). Online evaluation of coreference resolution. In *Proceedings of LREC 2004*, pages 1507–1510, Lisbon.

Popescu-Belis, A., Robba, I., and Sabah, G. (1998). Reference resolution beyond coreference: a conceptual frame and its application. In *Proceedings of COLING-ACL 1998*, pages 1046–1052, Montreal, Canada.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007a). OntoNotes: A unified relational semantic representation. In *Proceedings of ICSC 2007*, pages 517–526, Irvine, California.

Pradhan, S. S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007b). Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of ICSC 2007*, pages 446–453, Irvine, California.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Quinlan, R. (1993). *C4.5: Program for Machine Learning*. Morgan Kaufmann, San Francisco, California.

Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of EMNLP 2009*, pages 968–977, Suntec, Singapore.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Recasens, M. (2008). Discourse deixis and coreference: Evidence from AnCora. In *Proceedings of the Second Workshop on Anaphora Resolution (WAR II)*, volume 2 of *NEALT Proceedings Series*, pages 73–82, Bergen, Norway.

Recasens, M. (2009). A chain-starting classifier of definite NPs in Spanish. In *Proceedings of the EACL Student Research Workshop (EACL 2009)*, pages 46–53. Athens.

Recasens, M. and Hovy, E. (2009). A deeper look into features for coreference resolution. In Devi, S. L., Branco, A., and Mitkov, R., editors, *Anaphora Processing and Applications (DAARC 2009)*, volume 5847 of *LNAI*, pages 29–42, Berlin. Springer-Verlag.

Recasens, M. and Hovy, E. (2010). Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of ACL 2010*, pages 1423–1432, Uppsala, Sweden.

Recasens, M. and Hovy, E. (To appear). BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*.

Recasens, M., Hovy, E., and Martí, M. A. (2010a). A typology of near-identity relations for coreference (NIDENT). In *Proceedings of LREC 2010*, pages 149–156, Valletta, Malta.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010b). SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 1–8, Uppsala, Sweden.

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

Recasens, M., Martí, M. A., Taulé, M., Màrquez, L., and Sapena, E. (2009a). SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 70–75, Boulder, Colorado.

Recasens, M., Martí, M. A., and Taulé, M. (2009b). First-mention definites: More than exceptional cases. In Featherston, S. and Winkler, S., editors, *The Fruits of Empirical Linguistics*, volume 2, pages 217–237. Mouton de Gruyter, Berlin.

Rich, E. and LuperFoy, S. (1988). An architecture for anaphora resolution. In *Proceedings of ANLP 1988*, pages 18–24. Austin, Texas.

Rodríguez, K. J., Delogu, F., Versley, Y., Stemle, E., and Poesio, M. (2010). Anaphoric annotation of Wikipedia and blogs in the Live Memories Corpus. In *Proceedings of LREC 2010*, pages 157–163, Valletta, Malta.

Ruppenhofer, J., Sporleder, C., and Morante, R. (2010). SemEval-2010 Task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 45–50, Uppsala, Sweden.

Russell, B. (1905). On denoting. *Mind*, 15:479–493.

Sapena, E., Padró, L., and Turmo, J. (2010). RelaxCor: A global relaxation labeling approach to coreference resolution for the Semeval-2010 Coreference Task. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 88–91, Uppsala, Sweden.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT Workshop*, pages 47–50, Dublin.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, pages 777–784, Manchester, UK.

Shinyama, Y. and Sekine, S. (2003). Paraphrase acquisition for information extraction. In *Proceedings of the ACL 2nd International Workshop on Paraphrasing (IWP 2003)*, pages 65–71, Sapporo, Japan.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill, New York, second edition. Chapter 9.8.

Solà, J., editor (2002). *Gramàtica del català contemporani*. Empúries, Barcelona.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Steinberger, J., Poesio, M., Kabadjov, M. A., and Jeek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management: an International Journal*, 43(6):1663–1680.

Stevenson, R., Crawley, R., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9:519–548.

Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D., and Hysom, D. (2010). Coreference resolution with Reconcile. In *Proceedings of ACL 2010 Short Papers*, pages 156–161, Uppsala, Sweden.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of ACL-IJCNLP 2009*, pages 656–664, Suntec, Singapore.

Strube, M. and Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL 2003*, pages 168–175, Sapporo, Japan.

Strube, M., Rapp, S., and Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of ACL-EMNLP 2002*, pages 312–319.

Taboada, M. (2008). Reference, centers and transitions in spoken Spanish. In Gundel, J. and Hedberg, N., editors, *Reference and Reference Processing*, pages 176–215. Oxford University Press, Oxford.

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC 2008*, pages 96–101, Marrakech, Morocco.

Tetreault, J. (1999). Analysis of syntax-based pronoun resolution methods. In *Proceedings of ACL 1999*, pages 602–605, College Park, Maryland.

Tetreault, J. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL 2003*, pages 142–147. Edmonton, Canada.

Uryupina, O. (2003). High-precision identification of discourse-new and unique noun phrases. In *Proceedings of the ACL 2003 Student Workshop*, pages 80–86, Sapporo, Japan.

Uryupina, O. (2004). Linguistically motivated sample selection for coreference resolution. In *Proceedings of DAARC 2004*, Azores, Portugal.

Uryupina, O. (2006). Coreference resolution with and without linguistic knowledge. In *Proceedings of LREC 2006*, pages 893–898, Genoa, Italy.

Uryupina, O. (2007). *Knowledge Acquisition for Coreference Resolution*. PhD thesis, Saarland University, Saarbrücken, Germany.

Uryupina, O. (2008). Error analysis for learning-based coreference resolution. In *Proceedings of LREC 2008*, pages 1914–1919, Marrakech, Morocco.

Uryupina, O. (2010). Corry: A system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 100–103, Uppsala, Sweden.

van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

van Noord, G., Schuurman, I., and Vandeghinste, V. (2006). Syntactic annotation of large corpora in STEVIN. In *Proceedings of LREC 2006*, pages 1811–1814, Genoa, Italy.

Versley, Y. (2007). Antecedent selection techniques for high-recall coreference resolution. In *Proceedings of EMNLP-CoNLL 2007*, pages 496–505, Prague.

Versley, Y. (2008). Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6:333–353.

Versley, Y., Ponzetto, S., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008). BART: A modular toolkit for coreference resolution. In *Proceedings of LREC 2008*, pages 962–965, Marrakech, Morocco.

Vicedo, J. L. and Ferrández, A. (2006). Coreference in Q&A. In Strzalkowski, T. and Harabagiu, S., editors, *Advances in Open Domain Question Answering*, volume 32 of *Text, Speech and Language Technology*, pages 71–96. Springer-Verlag, Berlin.

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Vila, M., González, S., Martí, M. A., Llisterri, J., and Machuca, M. J. (2010). ClInt: a bilingual Spanish-Catalan spoken corpus of clinical interviews. *Procesamiento del Lenguaje Natural*, 45:105–111.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, pages 45–52.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of ICML 2009*, pages 577–584, Montreal, Canada.

Webber, B. L. (1979). *A Formal Approach to Discourse Anaphora*. Garland Press, New York.

Webber, B. L. (1988). Discourse deixis: reference to discourse segments. In *Proceedings of ACL 1988*, pages 113–122, Buffalo, New York.

Wick, M., Culotta, A., Rohanimanesh, K., and McCallum, A. (2009). An entity based model for coreference resolution. In *Proceedings of SDM 2009*, pages 365–376, Reno, Nevada.

Wick, M. and McCallum, A. (2009). Advances in learning and inference for partition-wise models of coreference resolution. Technical Report UM-CS-2009-028, University of Massachusetts, Amherst, Massachusetts.

Wintner, S. (2009). What science underlies Natural Language Engineering? *Computational Linguistics*, 35(4):641–644.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell, Oxford.

Yang, X. and Su, J. (2007). Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of ACL 2007*, pages 525–535, Prague.

Yang, X., Su, J., Lang, J., Tan, C. L., Liu, T., and Li, S. (2008). An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-HLT 2008*, pages 843–851, Columbus, Ohio.

Yang, X., Su, J., Zhou, G., and Tan, C. L. (2004). Improving pronoun resolution by incorporating coreferential information of candidates. In *Proceedings of ACL 2004*, pages 127–134, Barcelona.

Yang, X., Zhou, G., Su, J., and Tan, C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of ACL 2003*, pages 176–183, Sapporo, Japan.

Zaenen, A. (2006). Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.

Zhekova, D. and Kübler, S. (2010). UBIU: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 96–99, Uppsala, Sweden.

APPENDICES

APPENDIX A

System Outputs

This appendix provides, for two OntoNotes documents, the coreference outputs produced by different runs of CISTELL (i.e., STRONG MATCH, SUPER STRONG MATCH, BEST MATCH, WEAK MATCH; as explained in Section 4.3.2) as well as the ALL-SINGLETONS and HEAD-MATCH baselines. In addition, the outputs produced by the six systems that participated in SemEval (Section 6.4) are included for the first file. Coreferent mentions share the same subscript number.

A.1 OntoNotes file 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. **GOLD**

   [[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. **ALL-SINGLETONS BASELINE**

   [[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. **HEAD-MATCH BASELINE**

   [[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. **STRONG MATCH**

[[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 have 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. **Super strong match**

   [[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. **BEST MATCH**
[[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. **WEAK MATCH**
[[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 and 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 and 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 hear

[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]$_{9}$ de-

cides [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 OntoNotes file 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. **GOLD**
   [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. **ALL-SINGLETONS BASELINE**
   [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. **HEAD-MATCH BASELINE**
   [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. **STRONG MATCH**
   [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. **SUPER STRONG MATCH**
   [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. **Best match**
   [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. **Weak match**
   [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.

APPENDIX B

Near-identity Excerpts

This appendix includes the corpus of 60 excerpts that were used in groups of 20 in the three experiments described in Section 8.5. The excerpts were extracted from three electronic corpora—ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007a) and AnCora (Recasens and Martí, 2010)—as well as from the Web, a television show, and real conversation.

The task required coders to classify the selected pairs of NPs in each excerpt according to the (near-)identity relation(s) that obtained between them (Section 8.4). They had to assign one or more, but at least one, class to each pair of NPs. The answers are summarized in Appendix 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.

APPENDIX C

Answers to the Near-identity Task

This appendix reports the answers given by the six annotators in the near-identity task of Section 8.5 based on the excerpts included in Appendix B. The task required coders to classify the selected pairs of NPs in each excerpt according to the (near-)identity relation(s) that obtained between them (Section 8.4).

In the following three tables, the first column shows the excerpt number in parentheses and the ID numbers of the two NPs whose relation is under analysis. The rest of columns show the number of times a near-identity type (see the key at the beginning of each section) was assigned to each pair of NPs, only explicitly stated when different from 0. Note that rows summing up greater than six are cases for which one or more coders gave multiple answers.

C.1 Experiment 1

KEY

1 Non-identity; **2** Identity; **3** Near-identity;

3A Role; **3B** Location·Agency; **3C** Product·Producer; **3D** Informational realization;

3E Numerical function; **3F** Representation; **3Ga** Meronymy–Part·Whole; **3Gb** Meronymy–Set·Members;

3Gc Meronymy–Portion·Mass; **3Gd** Meronymy–Stuff·Object; **3Ha** Interpretation–Selection;

3Hb Interpretation–Viewpoint; **3Ia** Class–More specific; **3Ib** Class–More general;

3Ja Spatio-temporal func.–Place; **3Jb** Spatio-temporal func.–Time

Pair	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															

continued on next page

continued from previous page

Pair	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

Table C.1: Coders’ answers to Experiment 1

C.2 Experiment 2

KEY

1 Non-identity; **2** Identity; **3** Near-identity; **3Aa** Metonymy–Role; **3Ab** Metonymy–Location; **3Ac** Metonymy–Organization; **3Ad** Metonymy–Informational realization; **3Ae** Metonymy–Representation; **3Af** Metonymy–Other; **3Ba** Meronymy–Part·Whole; **3Bb** Meronymy–Set·Members; **3Bc** Meronymy–Stuff·Object; **3Bd** Meronymy–Overlap; **3Ca** Class–More specific; **3Cb** Class–More general; **3Da** Spatio-temporal func.–Place; **3Db** Spatio-temporal func.–Time; **3Dc** Spatio-temporal func.–Numerical func.; **3Dd** Spatio-temporal func.–Role func.

Pair	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								

Table C.2: Coders' answers to Experiment 2

C.3 Experiment 3

KEY

1 Non-identity; **2** Identity; **3** Near-identity;

3Aa Metonymy–Role; **3Ab** Metonymy–Location; **3Ac** Metonymy–Organization;

3Ad Metonymy–Informational realization; **3Ae** Metonymy–Representation; **3Af** Metonymy–Other;

3Ba Meronymy–Part·Whole; **3Bb** Meronymy–Stuff·Object; **3Bc** Meronymy–Overlap;

3Ca Class–More specific; **3Cb** Class–More general; **3Da** Spatio-temporal func.–Place;

3Db Spatio-temporal func.–Time; **3Dc** Spatio-temporal func.–Numerical func.;

3Dd Spatio-temporal func.–Role func.

Pair	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

Table C.3: Coders' answers to Experiment 3