Published in *Language Resources and Evaluation*. DOI: 10.1007/s10579-012-9194-z
The final publication is available at http://www.springerlink.com/content/006123u1h7812455

Coreference Resolution: An Empirical Study Based on SemEval-2010 Shared Task 1

Lluís Màrquez · Marta Recasens · Emili Sapena

Abstract This paper presents an empirical evaluation of coreference resolution that covers several interrelated dimensions. The main goal is to complete the comparative analysis from the SemEval-2010 task on *Coreference Resolution in Multiple Languages*. To do so, the study restricts the number of languages and systems involved, but extends and deepens the analysis of the system outputs, including a more qualitative discussion. The paper compares three automatic coreference resolution systems for three languages (English, Catalan and Spanish) in four evaluation settings, and using four evaluation measures. Given that our main goal is not to provide a comparison between resolution algorithms, these are merely used as tools to shed light on the different conditions under which coreference resolution is evaluated. Although the dimensions are strongly interdependent, making it very difficult to extract general principles, the study reveals a series of interesting issues in relation to coreference resolution: the portability of systems across languages, the influence of the type and quality of input annotations, and the behavior of the scoring measures.

Keywords Coreference resolution and evaluation · NLP system analysis · Machine learning based NLP tools · SemEval–2010 (Task 1) · Discourse entities

1 Introduction

Coreference resolution is the problem of identifying the expressions (usually NPs) in a text that refer to the same discourse entity. Despite the extensive work on this topic

L. Màrquez
TALP Research Center, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. Jordi Girona 1-3, 08034 Barcelona, Spain. E-mail: lluism@lsi.upc.edu

M. Recasens
CLiC Research Center, Departament de Lingüística, Universitat de Barcelona. Gran Via 585, 08007 Barcelona, Spain. E-mail: mrecasens@ub.edu

E. Sapena
TALP Research Center, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. Jordi Girona 1-3, 08034 Barcelona, Spain. E-mail: esapena@lsi.upc.edu

over the last years, it is still a highly challenging task in Natural Language Processing (NLP). Given a text like (1), the coreference community aims to build systems that automatically output "Major League Baseball," "its," and "the league" as mentions of the same entity, "its head of security" as a mention of a separate entity, and so forth.

(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.

A **discourse entity** (henceforth, entity) is defined as the collection of textual references to the same object in the discourse model, and each of these textual references is called a **mention**. Mentions of the same entity are said to **corefer**, whereas an entity that has one single mention is called a **singleton**. The terms "coreference" and "anaphora" are sometimes used interchangeably, but they are not always the same. A coreferent expression is only **anaphoric** if its interpretation depends on a previous expression in the text (i.e., its **antecedent**). In (1) above, *its* and *the league* are anaphoric, as the reader goes back in the text to find their antecedent. In contrast, a further mention of *Major League Baseball* using a lexical repetition would be coreferent but not anaphoric, as it could stand on its own.

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. Resolving these dependencies is necessary for discourse comprehension, and thus for NLP. However, the underlying process of how this is done is yet unclear, which makes the task of coreference resolution a real challenge. The mere task of producing the same results as those produced by humans is difficult and largely unsolved. There is nonetheless a strong interest in automatically identifying coreference links as they are needed by information extraction to merge different pieces of information referring to the same entity (McCarthy and Lehnert, 1995), by text summarization to produce a coherent and fluent summary (Azzam et al, 1999; Steinberger et al, 2007), by question answering to disambiguate references along a document (Morton, 1999; Vicedo and Ferrández, 2006), and by machine translation to translate pronouns correctly. Recently, state-of-the-art coreference resolution systems have been helpful for 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 arguments (Gerber and Chai, 2010; Ruppenhofer et al, 2010).

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). More recently, a task on Multilingual Coreference Resolution was organized at the SemEval-2010 evaluation exercise (Recasens et al, 2010). The goal of this task was to evaluate and compare automatic coreference resolution systems for six different languages in four evaluation settings and using four different evaluation measures. 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 mea-

sures. The task attracted considerable attention from a number of researchers, but only six teams submitted results. Moreover, participating systems did not run their systems for all the languages and evaluation settings, thus making direct comparisons among all the involved dimensions very difficult.

As discussed in the task description paper and slides,[1] the task contributed to the coreference community with valuable resources, evaluation benchmarks, and results along several dimensions. However, some problems were also identified and discussed. These were mainly related to the high complexity of the task, the limited number of participants, and a wrong design decision that did not allow a fair comparison between the settings using gold-standard input information and those using automatically predicted input information.

The current study shares the same fundamental motivations as SemEval-2010 Task 1, but places greater emphasis on analyzing the different conditions under which coreference resolution is evaluated rather than comparing different resolution algorithms. We provide a more thorough empirical analysis overcoming the aforementioned problems in the definition of the task. More precisely, greater insight is provided into: (1) coreference annotations across corpora and languages, (2) the evaluation measures and their different focus on assessing the quality of a system output, and (3) a qualitative analysis of the results, including commented examples.

To conduct such an in-depth analysis and keep every piece under control, some simplifications with respect to SemEval-2010 Task 1 were necessary. More specifically, we reduced the number of languages from six to three (English, Spanish, and Catalan), and we did not maintain the distinction between *closed* and *open* scenarios. Since this meant reevaluating the systems, we additionally restricted the comparison to three coreference resolution systems. Two of them, CISTELL and RELAXCOR, are in-house systems. The third one, RECONCILE, is freely available as open-source software.

With the aim of promoting continued research on this problem and the use of our data sets by the coreference community, we have made available all the corpora used in this study (i.e., the SemEval-2010 corpora updated with a few additional annotations), the scoring software upgraded with some new functionalities, and the system outputs of the different evaluation scenarios. For the latter we provide not only the regular textual representation, but also an HTML representation that can be viewed in any browser and where colors and meta-annotations facilitate the interpretation and comparison of the coreference annotations made by the three systems.[2] This is an additional contribution of this work, and it can also be used to reproduce the results reported here.

The rest of the paper is organized as follows. Section 2 presents the corpora used in this study, together with some statistics and an analysis of their main properties. Section 3 describes the three automatic systems for coreference resolution that are used in the study. Section 4 is devoted to the experimental setting, with a special emphasis on the evaluation measures. Section 5 provides the numerical results of the baselines and systems across languages and settings. In Section 6, a deeper analysis

[1] Available at the SemEval-2010 Task 1 website: http://stel.ub.edu/semeval2010-coref

[2] This material is available at http://nlp.lsi.upc.edu/coreference/LRE-2011/

Table 1 Size of the English, Catalan and Spanish corpora. The reported figures include the number of documents, sentences and lexical tokens for the training, development and test partitions

	Training			Development			Test		
	#docs	#sents	#tokens	#docs	#sents	#tokens	#docs	#sents	#tokens
English	229	3,648	79,060	39	741	17,044	85	1,141	24,206
Catalan	829	8,709	253,513	142	1,445	42,072	167	1,698	49,260
Spanish	875	9,022	284,179	140	1,419	44,460	168	1,705	51,040

of the system outputs is performed by focusing on more qualitative aspects and discussing specific examples. Finally, Section 7 concludes and identifies key issues for future research.

2 Corpora and Coreference Annotation

The corpora used in this study comprise the English, Catalan, and Spanish data sets from the SemEval-2010 Task 1 on Multilingual Coreference Resolution. These corpora are excerpts from the OntoNotes Release 2.0 (Pradhan et al, 2007) and AnCora corpora (Recasens and Martí, 2010). They contain coreference annotations of entities composed of pronouns and full noun phrases (including named entities), plus several annotation layers of syntactic and semantic information: lemma, part-of-speech, morphological features, dependency parsing, named entities, predicates, and semantic roles. Most of these annotation layers are doubly provided, once as *gold standard* and once as *predicted*, i.e., manually annotated versus predicted by automatic linguistic analyzers. The coreference annotation also includes the entities consisting of a single mention (singletons). For more information on these corpora, including formatting details and the linguistic processors used to produce the predicted layers of information, we refer the reader to the task description paper and website (Recasens et al, 2010).

The corpora of the three languages are divided into training, development and test sets following the same partitions as SemEval-2010 Task 1. The development corpora were used for parameter tuning. All the results reported in Sections 4, 5 and 6 were obtained on the test sets. Table 1 summarizes the number of documents (#docs), sentences (#sents), and tokens in the training, development and test sets. As can be seen, the Catalan and Spanish corpora are comparable in size, although the latter is slightly larger, while the English corpus is significantly smaller (about 30% of the total number of tokens).

Table 2 presents general statistics on the coreference annotation of the three corpora, which give a first impression of the similarities and differences between the languages. The first two blocks of rows show the absolute number (also averaged per document) of tokens, entities, mentions, and singletons.[3] It can be observed that the concentration of entities per document is larger in English, with an average of 68 entities per document, as opposed to 51 and 55 entities per document in Catalan and Spanish, respectively. This difference is largely explained by the high number

[3] The average number of entities per document is calculated as the summation of coreference chains in every document divided by the number of documents.

Table 2 Statistics on the coreference annotation for the English, Catalan and Spanish corpora

	English		**Catalan**		**Spanish**	
	#cases	avg.×doc.	#cases	avg.×doc.	#cases	avg.×doc.
Tokens	119,957	339.82	344,845	303.03	379,679	320.95
Mentions	32,943	93.32	94,447	82.99	105,144	88.88
Entities	24,018	68.04	58,169	51.11	65,007	54.95
Non-sing. Entities	3,302	9.35	14,253	12.52	15,177	12.83
Singletons	20,716	58.69	43,916	38.59	49,830	42.12
	62.88% of mentions		46.50% of mentions		47.39% of mentions	
	86.25% of entities		75.50% of entities		76.65% of entities	
	#mentions (excluding singletons)					
Avg. entity size	1.37 (3.70)		1.62 (3.55)		1.62 (3.64)	
	#tok	#sent #ment	#tok	#sent #ment	#tok	#sent #ment
Avg. distance to preceding mention	63.76	2.75 16.37	56.08	1.77 14.30	50.52	1.46 13.04
Decayed Density (Daumé and Marcu, 2005)	0.19		0.24		0.26	

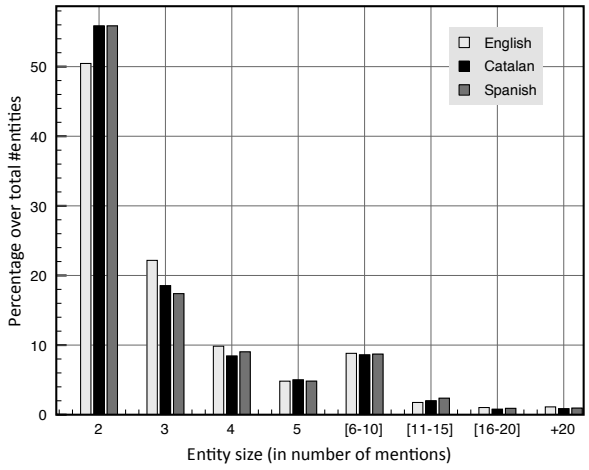

Fig. 1 Distribution (by language) of non-singleton entities according to entity size

of singleton entities in English, which represent ∼86% of the entities and ∼63% of the mentions. The same figures for Catalan and Spanish are about 15 and 10 points below, respectively.

The average entity size (in number of mentions) of the three languages is quite comparable if singletons are not taken into account. Unlike AnCora, where non-referential NPs like nominal predicates were filtered out, the OntoNotes corpus was only annotated with multi-mention entities. To make all the data sets as similar as possible for the SemEval shared task, singletons were identified heuristically in the English data set, although a few non-referential NPs that could not be automatically detected were unavoidably annotated as singletons. This accounts for the larger number of singletons in English. The bar chart in Figure 1 compares the distribution of non-

singleton entities according to entity size across the three languages. Remarkably, most of the cases (>50% of the total number of entities) fall into entities of size two. The distribution is very similar for all the languages. Catalan and Spanish show an almost identical distribution. English contains a slightly smaller number of entities of size two, but a slightly larger number of entities of size three (the variation is around five percent points).

The last two blocks of rows in Table 2 analyze how far apart coreferent mentions are from one another. To this end, we calculated the average distance to the preceding coreferent mention, and then averaged it on the entire corpus.[4] Three measurements of this distance were calculated by counting the number of word tokens (#tok), sentence boundaries (#sent), and mentions (#ment) in between two coreferent mentions. Additionally, this comparison was also measured in terms of *decayed density* (Daumé and Marcu, 2005). This measure was designed to capture the fact that some entities are referred to consistently across a document, while others are mentioned in only one short segment. This is a density measure, so it should correlate negatively with the distance-based measures. A clear picture emerges from this comparison: the English corpus contains a smaller number of dense entities, with mentions spread across longer distances on average and with more intervening mentions. Catalan and Spanish follow in this order. The fact that relative pronouns are annotated in the Catalan and Spanish data, but not in the English data, helps to account for the smaller distance between mentions observed in the first two languages.

We also carried out a more detailed study of the coreference relations in the three corpora. We grouped mentions into meaningful classes according to their morphology and their relation with the other mentions in the same coreference chain. The list of classes is described in Figure 2. They follow the ideas from Stoyanov et al (2009), but are adapted to our setting and languages. Given that Catalan and Spanish pronouns are always gendered, the P_3U class makes no sense for them. In the case of English, we omit the P_ELL and P_REL classes as it is not a pro-drop language, and relative pronouns are not coreferentially annotated in the English data set. Note that the same mention classes are used again in Section 6 to compute detailed results of the three coreference resolution systems.

Table 3 shows the number and percentage of each mention class for the three corpora. As can be seen, Catalan and Spanish present again a very similar distribution. When English is compared to the two Romance languages, we clearly observe that English has a higher number of PN_E coreference relations, but a lower number of CN_N. This can be accounted for by the same reason pointed out by Lundquist (2007) for Danish (of the same language family as English) and French (of the same language family as Catalan and Spanish). She observes a preference in French for *unfaithful* anaphors (that is, coreferent NPs whose head is different from that of the previous NP in the chain) that contrasts with the preference in Danish for *faithful* anaphors (that is, coreferent NPs that are pronouns or that repeat the previous coreferent NP). She attributes this difference to the different lexicalization patterns of Romance and Germanic languages. The former tend to lexicalize nouns at a more concrete and subordinate level, whereas the latter lexicalize more semantic

[4] Singletons are excluded.

Short Name	Description
PN_E	NPs headed by a Proper Name that match Exactly (excluding case and the determiner) at least one preceding mention in the same coreference chain
PN_P	NPs headed by a Proper Name that match Partially (i.e., head match or overlap, excluding case) at least one preceding mention in the same coreference chain
PN_N	NPs headed by a Proper Name that do **n**ot match any preceding mention in the same coreference chain
CN_E	Same definitions as in PN_E, PN_P and PN_N,
CN_P	but referring to NPs headed by a Common Noun
CN_N	
P_1∪2	First- and second-person pronouns that corefer with a preceding mention
P_3G	Gendered third-person pronouns that corefer with a preceding mention
P_3U	Ungendered third-person pronouns that corefer with a preceding mention
P_ELL	Elliptical pronominal subjects that corefer with a preceding mention
P_REL	Relative pronouns that corefer with a preceding mention

Fig. 2 Description of the mention classes considered in this study

Table 3 Number and percentage of coreference relations by mention class and language

	English		**Catalan**		**Spanish**	
	#cases	percent	#cases	percent	#cases	percent
PN_E	1,619	18.14%	4,282	11.80%	4,825	12.02%
PN_P	404	4.53%	566	1.56%	880	2.19%
PN_N	925	10.36%	2,210	6.09%	2,654	6.61%
CN_E	653	7.32%	4,141	11.41%	4,229	10.53%
CN_P	724	8.11%	4,014	11.06%	3,761	9.37%
CN_N	1,304	14.61%	7,990	22.02%	9,229	22.99%
P_1∪2	754	8.45%	353	0.97%	511	1.27%
P_3G	1,049	11.75%	2,239	6.17%	1,827	4.55%
P_3U	1,493	16.73%				
P_ELL			5,336	14.70%	6,856	17.08%
P_REL			5,147	14.18%	5,365	13.36%

features in verbs. As a result, Romance languages are said to be exocentric because they distribute the information onto the noun, and Germanic languages are said to be endocentric because they concentrate the information in the verb.

Also, if we assume that coreferent mentions that match exactly (i.e., PN_E and CN_E) are generally easier to resolve than non-matching mentions (PN_N and CN_N), we find that English shows a more favorable proportion of "easy" and "difficult" non-pronominal mention classes (25.46% – 24.97%) as compared to Catalan (23.21% – 28.11%) and Spanish (22.55% – 29.6%). This could certainly influence the relative performance across languages achieved by coreference resolution systems. In regard to pronouns, the small numbers of P_1∪2 and P_3G in Catalan and Spanish are counterbalanced by the large number of P_ELL, but English still has more pronouns altogether (∼37% versus ∼22% in the two Romance languages, excluding relative pronouns). The "emptiness" of elliptical subjects adds to the difficulty of resolving coreference for Catalan and Spanish.[5]

[5] It must be noted that, in this study, there is no need to recognize elliptical pronouns neither in the *gold* nor in the *predicted* setting, since they appear as special lexical tokens in the Catalan and Spanish corpora. They were inserted during the manual syntactic annotation of the AnCora corpora (Civit and Martí, 2005).

Table 4 Basic properties and configurations of the three systems used in the evaluation. We differentiate between the *Classification* and *Linking* process even for those systems that do resolution in one step

Property	CISTELL	RELAXCOR	RECONCILE
Classification model	Entity-mention	Mention-pair	Mention-pair
Classification algorithm	TiMBL	Constraints from DT	Perceptron
Classification+Linking	One step	One step	Two steps
Linking algorithm	Agglomerative clustering	Relaxation labeling	Single link
Machine learning	Supervised	Supervised	Supervised
# Features	30–32	> 100	60
Use of WordNet	Yes	Yes	Yes
Training process	Train	Train and development	Train[a]
Optimized for English	No	Yes	Yes
Scenario-specific training	Yes	Only development	Yes
Languages	English, Catalan, Spanish	English, Catalan, Spanish	English

[a] The RECONCILE system offers the option of adjusting the coreference decision threshold on the development set, but we used the default value of 0.5.

3 Coreference Systems

This section introduces the three coreference systems that were used in the study. They represent the main classes of supervised learning coreference systems according to the classification model, and to the way the classification and linking steps are integrated. Classification models mainly fall into **mention-pair** and **entity-mention** models (Ng, 2010). The former classify every pair of mentions as coreferent or not. This is the model followed by RECONCILE and RELAXCOR. The latter models, used by CISTELL, define an entity as a bag of (ordered) mentions and extract a set of properties defining the whole entity, then classification is done by comparing mentions with entities. In terms of integrating classification and linking, a distinction can be drawn between **two-step** and **one-step** models. RECONCILE is a two-step system because it first classifies all the mention pairs, and then links the mentions to entities. In contrast, CISTELL and RELAXCOR are one-step systems because they collapse classification and linking into a single step.

In terms of features, the three systems use a similar feature set that captures the information classically used by coreference systems: textual strings (e.g., head match, substring match, distance), morphology (e.g., NP type, gender, number), syntax (e.g., grammatical function), and semantics (e.g., NE type, synonymy/hypernymy relations in WordNet). The difference in the size of the feature set, ranging from the 30 features used by CISTELL to the over 100 features used by RELAXCOR, generally stems from different choices in terms of binary or multi-valued features, rather than different kinds of information.

Table 4 reports the main properties and configurations of the three systems used in our study. The reader will find a detailed description of the aspects included in the table in the following subsections 3.1, 3.2 and 3.3, respectively devoted to CISTELL, RELAXCOR and RECONCILE.

3.1 CISTELL

The approach taken in devising the CISTELL coreference system (Recasens, 2010) adds to the body of work on entity-mention models. These models are meant to determine 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. Luo et al (2004) pioneered this line of research, and concluded that it is "an area that needs further research." CISTELL is based on the belief that 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, and to this end it handles discourse entities as (growing) baskets.[6] 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 discourse entity so that the information of subsequent references can be stored in it as the discourse progresses.

After identifying the set of mentions, CISTELL allocates to each mention a basket that contains mention attributes such as head, gender, number, part-of-speech, NE type, modifiers, grammatical role, synonyms, hypernyms, sentence position, etc. The convenient property of baskets is that they can *grow* by swallowing other baskets and incorporating their attributes. 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).

CISTELL follows the learning-based coreference architecture in which the task is split into classification and linking (Soon et al, 2001), but combines them simultaneously. A pairwise classifier that predicts the probability of two mentions coreferring is trained with the TiMBL memory-based learning software package (Daelemans et al, 1999). It 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. The 30 learning features that were used in this study for English, and the 32 learning features that were used for Catalan and Spanish, are a subset of those described in Recasens and Hovy (2009). Separate classifiers were trained for each of the evaluation scenarios, depending on whether the annotation was gold-standard or predicted, and whether true or system mentions were used.

Linking is identified with basket growing, the core process, that calls the pairwise classifier every time it considers whether a basket must be clustered into a (growing) basket. When the two baskets are singletons, they are linked if they are classified as coreferent by the classifier. Otherwise, the basket under analysis is paired with each of the baskets contained within the larger basket, and it is only linked if all the pairs are classified as coreferent. This is how the *strong match* model behaves, which turned out to obtain the best results among all the evaluated techniques for basket growing.

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

3.2 RELAXCOR

RELAXCOR (Sapena et al, 2010a) is a coreference resolution system based on constraint satisfaction. It represents the problem as a graph connecting any pair of candidate coreferent mentions, and it applies relaxation labeling over a set of constraints to decide the set of most compatible coreference relations. This approach combines classification and linking in one step. Thus, decisions are taken considering the entire set of mentions, which ensures consistency and avoids local classification decisions.

The knowledge of the system is a set of weighted constraints. Each constraint has an associated weight reflecting its confidence. The sign of the weight indicates whether a pair or group of mentions corefer (positive) or not (negative). Only constraints over pairs of mentions were used in the current version. However, RELAXCOR can handle higher-order constraints. Constraints can be obtained from any source, including a training data set from which they can be manually or automatically acquired. For the present study, all constraints were learned automatically using more than a hundred features over the mention pairs in the training sets. The typical attributes were used, like those in Sapena et al (2010b), but binarized for each possible value. In addition, other features that could help, such as whether a mention is an NE of location type or a possessive phrase, were included. A decision tree was generated from the training data set, and a set of constraints was extracted with the C4.5 rule-learning algorithm (Quinlan, 1993). The so learned constraints are conjunctions of attribute-value pairs. The weight associated with each constraint is the constraint precision minus a balance value, which is determined using the development step.

The coreference resolution problem is represented as a graph with mentions in the vertices. Mentions are connected to each other by edges. Edges are assigned a weight that indicates the confidence that the mention pair corefers or not. More specifically, an edge weight is the sum of the weights of the constraints that apply to that mention pair. The larger the edge weight in absolute terms, the more reliable.

RELAXCOR uses relaxation labeling for the resolution process. Relaxation labeling is an iterative algorithm that performs function optimization based on local information (Hummel and Zucker, 1987). It has been widely used to solve NLP problems such as part of speech tagging (Padró, 1998) and opinion mining (Popescu and Etzioni, 2005). An array of probability values is maintained for each vertex/mention. Each value corresponds to the probability that the mention belongs to a specific entity given all the possible entities in the document. During the resolution process, the probability arrays are updated according to the edge weights and probability arrays of the neighboring vertices. The larger the edge weight, the stronger the influence exerted by the neighboring probability array. The process stops when there are no more changes in the probability arrays or the maximum change does not exceed an *epsilon* parameter.

The RELAXCOR implementation used in the present study is an improved version of the system that participated in the SemEval-2010 Task 1 (Sapena et al, 2010b). The largest differences involve the training and development processes. The current RELAXCOR includes a parameter optimization process using the development data sets. The optimized parameters are *balance* and *pruning*. The former adjusts the constraint weights to improve the balance between precision and recall; the latter limits the num-

ber of neighbors that a vertex can have. Limiting the number of neighbors reduces the computational cost significantly and improves overall performance too. Optimizing this parameter depends on properties like document size and the quality of the information given by the constraints. Both parameters were empirically adjusted on the development set for the CEAF evaluation measure.

3.3 RECONCILE

In addition to CISTELL and RELAXCOR we decided to include a third system to gain a better insight into coreference resolution. There are only a few freely available coreference systems, such as BART (Versley et al, 2008), the Illinois Coreference Package (Bengtson and Roth, 2008), Reconcile (Stoyanov et al, 2010), and OpenNLP.[7] Given that we wanted the three systems of our study to solve coreference using the same input information, we needed a system that accepted an already preprocessed document as input. After reviewing the different options, we chose RECONCILE as it satisfied our needs with minimal effort. However, RECONCILE, as well as the rest of publicly available systems, only works for English. Indeed, there seems to exist no language-independent coreference system.

The RECONCILE system is different from CISTELL and RELAXCOR in that it is a platform meant as a research testbed that can be easily customized by the user to experiment with different coreference resolution architectures, learning algorithms, feature sets, data sets, and scoring measures. In this way, it facilitates consistent comparisons of different coreference resolution systems (for English). The structure of RECONCILE is best described by the seven desiderata that guided its design: (i) to implement the basic architecture of state-of-the-art learning-based coreference resolution systems; (ii) to support experimentation on the MUC and ACE data sets; (iii) to implement the most popular coreference resolution scoring measures; (iv) to create an end-to-end coreference resolver that achieves state-of-the-art performance (using its default configuration); (v) to make it easily extendable with new methods and features; (vi) to make it relatively fast and easy to configure and run; (vii) to include a set of pre-built resolvers that can be used as black-box coreference resolution systems.

The basic architecture of RECONCILE includes five major steps. Firstly, it preprocesses the data using a sentence splitter, tokenizer, POS tagger, parser, NER, and NP detector. Secondly, it produces feature vectors for every NP pair, including over 80 features inspired by Soon et al (2001) and Ng and Cardie (2002). Thirdly, it learns a classifier that assigns a score indicating the likelihood that a pair of NPs is coreferent. Fourthly, it employs clustering to form the final set of entities. Finally, it evaluates the output according to the MUC, B^3, and CEAF scores. For the experiment reported in this paper, we discarded the first and final steps, and we used the default configuration, namely the Reconcile_{2010} implementation, which includes a hand-selected subset of 60 features, an averaged perceptron classifier, and a single-link clustering with a positive decision threshold of 0.5.

Adapting the system to work in our experimental setting required only a minimal effort of format conversion at the input and output of the RECONCILE module. In

[7] http://opennlp.sourceforge.net

contrast, the effort necessary to port the system to Spanish and Catalan would have required substantial programming and extensive knowledge of the system implementation. This is why we report RECONCILE scores only for the English data set.

4 Experimental Setup

4.1 Evaluation Scenarios

Four different evaluation scenarios are considered in this work, differing along two dimensions: (1) *true* versus *system* mentions, and (2) *gold* versus *predicted* input information. Combining these two dimensions yields four different settings, which allow us to study the differences of solving coreference relations under an ideal scenario versus a more realistic one, in which mention boundaries and all the input linguistic features have to be automatically predicted.[8]

True mentions as well as gold and predicted morphosyntactic layers of annotation were already available in the SemEval-2010 Task 1 data sets (see Section 2 for more details), while system mentions were supposed to be generated by the participating systems. In this work, we implemented a simple mention detection procedure for supplying mentions so that the three coreference resolution systems use the same set of mentions in the *system mention* scenarios. System mentions are included as new annotation columns in the updated data sets released with this work.

The mention detection algorithm adds one mention for every noun and pronoun encountered in the text, except for multiple consecutive nouns (in this case, the mention is added for the last noun, a heuristic for identifying the syntactic head). Nouns and pronouns are detected by checking their part-of-speech tag. Mention boundaries are determined by looking at the dependency syntactic tree (either gold or predicted, depending on the setting) and selecting the complete segment of text that is covered by the noun or pronoun under analysis. That is, the rightmost (or leftmost) dependency modifying the noun is recursively followed to locate the right (or left) mention boundary. This simple mention detection algorithm can be considered as a baseline, but it performs reasonably well (especially for English), as reported in Section 5. Typical errors made by the system mention extraction procedure include extracted NPs that are not referential (e.g., predicative and appositive phrases), mentions with incorrect boundaries, and mentions that are not correctly extracted in a sequence of nouns (due to the NP head heuristic). Obviously, the number of errors increases with predicted annotations.

[8] The evaluation of SemEval-2010 Task 1 (Recasens et al, 2010) also distinguished between *closed* and *open* settings. In the former, systems had to be built strictly with the information provided in the task data sets. In the latter, systems could be developed using any external tools and resources (e.g., WordNet, Wikipedia, etc.). In this study we do not make such a distinction because the three systems rely on the same sources of information: training set, particular heuristics, and WordNet.

4.2 Evaluation Measures

Automatic evaluation measures are crucial for coreference system development and comparison. Unfortunately, there is no agreement at present on a standard measure for coreference resolution evaluation. This is why we included the three measures most widely used to assess the quality of a coreference output—namely B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005), and MUC (Vilain et al, 1995)—plus the recently developed BLANC (Recasens and Hovy, 2011), to provide a more complete picture of the behavior of the different evaluation approaches. B^3 and CEAF are mention-based, whereas MUC and BLANC are link-based.

The following describes in more detail what each measure quantifies as well as its strengths and weaknesses. In evaluating the output produced by a coreference resolution system, we need to compare the true set of entities (the **key** or **key partition**, i.e., the manually annotated entities) with the predicted set of entities (the **response** or **response partition**, i.e., the entities output by a system). Entities are viewed as sets of mentions. The **cardinality** of an entity is the number of mentions it contains. The mentions in the key are known as **true mentions**, and the mentions in the response are known as **system mentions**. The MUC, B^3 and CEAF results are expressed in terms of precision (P), recall (R), and F_1, which is defined as the harmonic mean between precision and recall as usual: $F_1 = 2 \cdot P \cdot R \,/\, (P + R)$.

4.2.1 The MUC scoring algorithm

The MUC scoring algorithm was first introduced by the MUC-6 evaluation campaign in 1995. It operates by comparing the entities defined by the links in the key and the response. In short, it counts the least number of links that need to be inserted in or deleted from the response to transform its entities into those of the key. The resulting formula (1) takes the set of entities in the key (to compute recall) or in the response (to compute precision) as S, and finds the partition of S, namely $p(S)$, relative to the response (to compute recall) or to the key (to compute precision). For instance, for each entity S_i, recall finds $p(S_i)$, i.e., the partition that results from intersecting S_i and those entities in the response that overlap S_i, including implicit singletons. Precision works the other way around and takes the response as S.

$$\begin{array}{l}\text{MUC Recall } (S \text{ is the key}) \\ \text{MUC Precision } (S \text{ is the response})\end{array} = \frac{\sum_{i=1}^{n}(|S_i| - |p(S_i)|)}{\sum_{i=1}^{n}(|S_i| - 1)} \qquad (1)$$

As observed by many (Bagga and Baldwin, 1998; Luo, 2005), the MUC measure is severely flawed for two main reasons. First, it is too lenient with entities containing wrong mentions: classifying one mention into a wrong entity counts as one precision and one recall error, while completely merging two entities counts as a single recall error. This can easily result in higher F-scores for worse systems. Finkel and Manning (2008) point out that if all the mentions in each document of the MUC test sets are linked to one single entity, the MUC measure gives a score higher than any published system. Second, given that it only takes into account coreference links, it ignores correct singleton entities. It is only when a singleton mention is incorrectly linked to another mention that precision decreases. For this reason, this measure is not a good

choice when working with data sets that, unlike the MUC corpora (Hirschman and Chinchor, 1997), are annotated with singletons.

4.2.2 B-CUBED (B^3)

The B^3 measure was developed in response to the shortcomings of MUC. It shifts the attention from links to mentions by computing precision and recall for each mention, and then taking the weighted average of these individual precision and recall scores. For a mention m_i, the individual precision represents how many mentions in the response entity of m_i corefer. The individual recall represents how many mentions in the key entity of m_i are output as coreferent. The formula for recall for a given mention m_i is given in (2), and that for precision is given in (3), where R_{m_i} is the response entity of mention m_i, and K_{m_i} is the key entity of mention m_i. Their cardinality is the number of mentions. The final precision and recall are computed by averaging these scores over all the mentions.

$$\text{B}^3 \text{ Recall}(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|K_{m_i}|} \tag{2}$$

$$\text{B}^3 \text{ Precision}(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|R_{m_i}|} \tag{3}$$

However, this measure has also been criticized. Luo (2005) considers that B^3 can give counterintuitive results due to the fact that an entity can be used more than once when computing the intersection of the key and response partitions. Besides, Recasens and Hovy (2011) point out another weakness. When working with corpora where all entities are annotated and singletons appear in large numbers, scores rapidly approach 100%. More seriously, outputting all the mentions as singletons obtains a score close to some state-of-the-art performances.

4.2.3 Constrained Entity-Alignment F-Measure (CEAF)

Luo (2005) proposed CEAF to solve the problem of reusing entities in B^3. It finds the best one-to-one mapping between the entities in the key and the response, i.e., each response entity is aligned with at most one key entity. The best alignment is the one maximizing the total entity similarity —denoted as $\Phi(g^*)$— and it is found by the Kuhn-Munkres algorithm. Two similarity functions for comparing two entities are suggested, resulting in the mention-based CEAF and the entity-based CEAF that use (4) and (5), respectively, where *K* refers again to the key partition, and *R* to the response partition.

$$\phi_3(K_i, R_i) = |K_i \cap R_i| \tag{4}$$

$$\phi_4(K_i, R_i) = \frac{2|K_i \cap R_i|}{|K_i| + |R_i|} \tag{5}$$

We use the mention-based CEAF to score the experiments reported in this paper because it is the most widely used. It corresponds to the number of common mentions between every two aligned entities divided by the total number of mentions. When

the key and response have the same number of mentions, recall and precision are the same. On the basis of the best alignment, they are computed according to (6) and (7).

$$\text{CEAF Recall} = \frac{\Phi(g^*)}{\sum_i \phi(K_i, K_i)} \tag{6}$$

$$\text{CEAF Precision} = \frac{\Phi(g^*)}{\sum_i \phi(R_i, R_i)} \tag{7}$$

Again, CEAF is not free of criticism. It suffers from the singleton problem just as B^3 does, which accounts for the fact that B^3 and CEAF usually get higher scores than MUC on corpora such as ACE where singletons are annotated, because a great percentage of the score is simply due to the resolution of singletons. In addition, the entity alignment of CEAF might cause a correct coreference link to be ignored if that entity finds no alignment in the key (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.

4.2.4 BiLateral Assessment of Noun-phrase Coreference (BLANC)

The main motivation behind the BLANC measure is to take the imbalance of singleton vs. coreferent mentions into account. To this end, it returns to the idea of links, but with a fundamental difference with respect to MUC: it considers the two aspects of the problem, namely not only coreference links but also non-coreference links (i.e., those that hold between every two mentions that do not corefer). The sum of the two remains constant across the key and response. Although this is an idea that comes from the Rand index (Rand, 1971), BLANC puts equal emphasis on each type of link by computing precision and recall separately for coreference and non-coreference links, and then averaging the two precision or recall scores for the final score. This is shown in (8) and (9), 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. Finally, the BLANC score averages the F-score for coreference links and the F-score for non-coreference links.

$$\text{BLANC Recall} = \frac{rc}{2(rc + wn)} + \frac{rn}{2(rn + wc)} \tag{8}$$

$$\text{BLANC Precision} = \frac{rc}{2(rc + wc)} + \frac{rn}{2(rn + wn)} \tag{9}$$

Four simple variations are defined for those cases when either the key or the response partition contains only singletons or a single entity. Unlike B^3 and CEAF, a coreference resolution system has to get high precision and recall for *both* coreference and non-coreference simultaneously to score well under BLANC. Although it is a very new measure and has not undergone extensive testing yet, its main weakness is revealed in the not very likely scenario of a document that consists of singletons except for one two-mention entity, as BLANC would penalize too severely a system that outputs all the mentions as singletons.

4.2.5 Evaluating on System Mentions

An issue that has been discussed by various authors (Bengtson and Roth, 2008; Stoyanov et al, 2009; Rahman and Ng, 2009; Cai and Strube, 2010) is the assumption made by B^3, CEAF and BLANC that the mention set in the key partition is the same as the mention set in the response partition. Arguably, end-to-end systems may output some mentions that do not map onto any true mention, or vice versa, some true mentions may not map onto any system mention. These are called **twinless** mentions by Stoyanov et al (2009). To handle twinless mentions, the above measures have been implemented with minor tweaks.

Bengtson and Roth (2008) simply discard twinless mentions, while Stoyanov et al (2009) suggest two variants of B^3: $B^3$0 and B^3all. The former discards twinless system mentions and sets recall(m_i) = 0 if m_i is a twinless true mention; the latter retains twinless system mentions, and sets precision(m_i) = $\frac{1}{|R_{m_i}|}$ if m_i is a twinless system mention, and recall(m_i) = $\frac{1}{|K_{m_i}|}$ if m_i is a twinless true mention. Another adjustment for both B^3 and CEAF is proposed by Rahman and Ng (2009): they remove only those twinless system mentions that are singletons, as they argue that in these cases the system should not be penalized for mentions that it has successfully identified as singletons. Recently, Cai and Strube (2010) have pointed out several outputs that are not properly evaluated by any of the above approaches. To deal with system mentions more successfully, they present two variants of B^3 and CEAF that (i) insert twinless true mentions into the response partition as singletons, (ii) remove twinless system mentions that are resolved as singletons, and (iii) insert twinless system mentions that are resolved as coreferent into the key partition (as singletons).

At a closer look, it appears that the two variants introduced by Cai and Strube (2010) can be regarded as adjustments of the key and response partitions rather than variants of the evaluation measures themselves. By adjusting the two partitions, each true mention can be aligned to a system mention, so that both the key and response partitions have the same number of mentions, and systems are neither unfairly favored nor unfairly penalized. We realized that the three adjustments by Cai and Strube (2010) for B^3 and CEAF make it possible to apply any coreference evaluation measure, and this is the approach followed in this paper to evaluate the *system mentions* $\times$ *gold annotation* and *system mentions* $\times$ *predicted annotation* scenarios. This new adjustment is a contribution that has been already incorporated into the scoring software. This software, which is distributed with the rest of materials of the paper, has also been adopted by the CoNLL-2011 shared task (Pradhan et al, 2011) as the official scorer.

4.2.6 Evaluating Mention Detection

Performance on the task of mention detection alone is measured in Table 6 with recall, precision, and F_1. System mentions are rewarded with 1 point if their boundaries coincide with those of the true mentions, with 0.5 points if their boundaries are within the true mention including its head, and with 0 otherwise.

5 Baseline and System Results

This section presents the results of the CISTELL, RELAXCOR and RECONCILE coreference resolution systems on the SemEval data. Before this, Tables 5 and 6 provide relevant information to interpret the results of the three systems. Table 5 shows the scores of two naive baselines together with oracle scores, and Table 6 shows the results of CISTELL, RELAXCOR and RECONCILE on the mention detection task.

5.1 Baseline Scores

The two baselines reported in Table 5 represent the most straightforward outputs: (i) SINGLETONS does not create any coreference link, but considers each mention as a separate entity, and (ii) ALL-IN-ONE groups all the document mentions into one single entity. The ORACLE represents the best results achievable given a particular mention detection setting. Obviously, 100% for the four evaluation measures is only achievable when *true mentions* are used.

We only provide the SINGLETONS scores once for each language as using *true mentions* or *system mentions* does not make any difference in the final score if no coreference link is output. This is so, however, due to the adjustment of the outputs that we make inspired by Cai and Strube (2010). As explained above in Section 4.2, twinless true mentions are inserted into the response partition as singletons, and singleton twinless system mentions are removed. This invariance is evidence that Cai and Strube's (2010) adjustment makes it possible for the coreference resolution measures to strictly evaluate coreference resolution without being influenced by mention detection performance.

Surprisingly enough, the ALL-IN-ONE baseline using system mentions obtains higher scores than the one using true mentions according to CEAF, B^3 and BLANC in the three languages. The fact that only MUC behaves as initially expected hints at the most plausible explanation: the difference is due to singletons as well as to Cai and Strube's (2010) adjustment for aligning true and system mentions. Unavoidably, a large number of true mentions are missing from the set of system mentions, but the adjustment inserts them into the response partition as singletons, thus they are not included into the same entity as all the mentions automatically detected. If we also keep in mind that the majority of mentions are singletons, especially long and syntactically complex NPs that are hard to detect automatically, twinless true mentions that escape from being included in the ALL-IN-ONE entity account for the increase in performance.

These simple baselines reveal limitations of the evaluation measures on the two extremes (see Section 4.2): CEAF and B^3 reward the naive SINGLETONS baseline too much, while MUC gives a too high score to the naive ALL-IN-ONE baseline. As a result, Table 5 also illustrates differences between the data sets. The English data set obtains the highest CEAF and B^3 scores for the SINGLETONS baseline, whereas the Catalan and Spanish data sets obtain the highest MUC scores for the ALL-IN-ONE baseline. This is easily accounted for by the slightly larger number of singletons in

Table 5 Baseline and *oracle* scores across all settings, languages and evaluation measures. SINGLETONS: Each mention forms a separate entity. ALL-IN-ONE: All mentions are grouped into one single entity. ORACLE: Best results achievable given a particular mention detection setting

	CEAF	MUC			B^3			BLANC		
	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
English										
SINGLETONS	71.2	0.0	0.0	0.0	71.2	100	83.2	50.0	49.2	49.6
true mentions										
ALL-IN-ONE	10.5	100	29.2	45.2	100	3.5	6.7	50.0	0.8	1.6
ORACLE	100	100	100	100	100	100	100	100	100	100
system mentions based on gold syntax										
ALL-IN-ONE	19.8	76.1	24.7	37.3	91.3	17.6	29.5	45.7	49.7	23.0
ORACLE	93.1	76.1	100	86.4	90.8	100	95.2	81.8	99.7	88.7
system mentions based on predicted syntax										
ALL-IN-ONE	23.0	72.7	23.6	35.7	90.6	21.4	34.6	47.5	49.9	26.7
ORACLE	92.1	72.7	100	84.2	89.5	100	94.5	80.2	99.7	87.5
Catalan										
SINGLETONS	61.2	0.0	0.0	0.0	61.2	100	75.9	50.0	48.7	49.3
true mentions										
ALL-IN-ONE	11.8	100	39.3	56.4	100	4.0	7.7	50.0	1.3	2.6
ORACLE	100	100	100	100	100	100	100	100	100	100
system mentions based on gold syntax										
ALL-IN-ONE	22.0	70.8	27.7	39.8	88.2	20.7	33.6	48.4	49.9	26.2
ORACLE	88.7	70.8	100	82.9	85.5	100	92.2	82.0	99.5	88.8
system mentions based on predicted syntax										
ALL-IN-ONE	24.9	60.1	23.1	33.4	85.5	26.1	39.9	47.2	49.8	31.1
ORACLE	84.5	60.1	100	75.1	80.6	100	89.2	75.8	99.3	83.7
Spanish										
SINGLETONS	62.2	0.0	0.0	0.0	62.2	100	76.7	50.0	48.8	49.4
true mentions										
ALL-IN-ONE	11.9	100	38.3	55.4	100	3.9	7.6	50.0	1.2	2.4
ORACLE	100	100	100	100	100	100	100	100	100	100
system mentions based on gold syntax										
ALL-IN-ONE	21.4	70.5	27.8	39.9	87.6	20.5	33.3	45.9	49.7	26.1
ORACLE	88.8	70.5	100	82.7	85.4	100	92.1	79.5	99.5	86.9
system mentions based on predicted syntax										
ALL-IN-ONE	25.5	59.2	23.3	33.5	84.9	27.1	41.1	46.2	49.8	31.7
ORACLE	84.6	59.2	100	74.4	80.4	100	89.1	74.2	99.4	82.3

the English data (see Section 2). Because of the 50% recall upper limit of BLANC, the SINGLETONS baseline scores considerably lower.

The ORACLE agrees with our expectations, except for the very small difference in English in the performance using system mentions based on gold syntax with respect to that based on predicted syntax, as opposed to the seven- or eight-point difference observed in Catalan and Spanish. There are two reasons for this. First, the English parser performs better than the Catalan and Spanish counterparts. Wrong PoS tag or dependency relations are likely to have a negative effect on the quality of mention detection. As shown in Table 6, the decrease in mention detection performance in Catalan and Spanish using predicted syntax is considerably larger than in English. Second, the smaller decrease in English may have to do with the fact that the mention detection architecture was originally designed for English and not particularly adapted for either Catalan/Spanish or the different data sets (see the description in Section 4.1).

Table 6 Mention detection results (Recall, Precision and F_1) for the three systems across all settings and languages

	English			Catalan			Spanish		
	R	P	F_1	R	P	F_1	R	P	F_1
true mentions × *predicted annotation*									
CISTELL	85.4	89.0	87.2	82.5	86.4	84.4	83.3	87.1	85.2
RELAXCOR	100	100	100	100	100	100	100	100	100
RECONCILE	100	100	100						
system mentions × *gold annotation*									
CISTELL	83.5	92.9	87.9	77.6	77.3	77.4	78.5	81.0	79.7
RELAXCOR	83.3	92.7	87.8	77.5	77.2	77.3	78.5	80.9	79.7
RECONCILE	83.0	92.7	87.6						
system mentions × *predicted annotation*									
CISTELL	75.8	84.3	79.9	65.1	63.8	64.5	65.0	66.9	66.0
RELAXCOR	75.8	84.3	79.8	65.1	63.8	64.5	65.0	66.9	66.0
RECONCILE	75.6	84.4	79.8						

5.2 Mention Detection Scores

Table 6 shows that, even if the quality of the mention detection module is high, especially for English, it represents a drop of 12 points with respect to true mentions (from 100% to ∼88%), and a further drop of 8 points when detection is based on predicted instead of gold annotation (from ∼88% to ∼80%). The results are between 10 and 15 points lower for Catalan and Spanish. These drops are not so sharp in the ORACLE (Table 5) because of the singleton adjustment for mapping the response onto the key partition that inserts missing singletons. Although our initial goal was to have the three systems use the same set of true mentions and system mentions, Table 6 shows that, unlike RELAXCOR and RECONCILE, CISTELL did not reach 100% in the *true mentions* × *predicted annotation* setting. Although true mentions were provided, CISTELL is highly dependent on the syntactic tree as it requires mentions to coincide with a syntactic node, which is clearly not always the case when predicted annotations are used.

5.3 System Scores

Table 7 displays the results of CISTELL and RELAXCOR for English, Catalan, Spanish, and averaged results for the three languages, as well as the results of RECONCILE for English, in the four evaluation settings and according to the four evaluation measures. Results are presented sequentially by language and setting. Unlike the corresponding table in the task description article of *Proceedings of SemEval-2010* (Recasens et al, 2010), all the cells of Table 7 are filled except for the Catalan and Spanish results of RECONCILE. We are then in a better position to compare coreference systems at multiple levels. This section presents the results from a quantitative point of view, while the next section tries to shed additional light on these results and provides some qualitative discussion.

Table 7 Results of the three systems across all languages, settings and evaluation measures

	CEAF	MUC			B^3			BLANC		
	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
English										
true mentions × gold annotation										
CISTELL	72.73	47.12	43.38	45.17	79.61	79.07	79.34	63.85	68.14	65.69
RELAXCOR	82.98	59.87	74.61	66.43	84.54	91.96	88.09	73.37	81.63	76.86
RECONCILE	77.16	30.27	76.84	43.43	76.44	96.70	85.39	60.05	83.24	65.24
true mentions × predicted annotation										
CISTELL	73.22	44.68	41.76	43.17	79.57	79.69	79.63	63.84	68.46	65.80
RELAXCOR	80.79	52.72	74.12	61.61	81.83	92.96	87.04	68.26	81.71	73.11
RECONCILE	75.99	26.74	75.77	39.54	75.46	96.97	84.87	56.91	81.10	61.06
system mentions × gold annotation										
CISTELL	71.84	38.15	40.77	39.42	77.83	81.95	79.84	59.23	66.71	61.83
RELAXCOR	78.70	45.20	62.15	52.34	80.38	89.83	84.84	64.41	73.05	67.69
RECONCILE	75.12	19.33	76.12	30.83	74.45	97.85	84.56	55.38	83.46	58.98
system mentions × predicted annotation										
CISTELL	72.00	37.42	40.35	38.83	78.20	82.04	80.07	59.62	66.11	62.00
RELAXCOR	77.47	36.02	63.70	46.02	78.07	92.77	84.79	61.18	78.24	65.89
RECONCILE	73.90	14.61	75.00	24.46	73.39	98.35	84.05	52.93	79.86	55.01
Catalan										
true mentions × gold annotation										
CISTELL	68.81	43.55	47.05	45.23	71.68	76.50	74.01	64.40	67.93	65.95
RELAXCOR	74.27	55.76	66.72	60.75	75.56	85.43	80.19	63.78	72.10	66.89
true mentions × predicted annotation										
CISTELL	67.47	37.58	41.19	39.30	71.34	76.43	73.79	60.91	64.86	62.55
RELAXCOR	74.26	55.72	67.93	61.23	75.25	86.63	80.54	62.06	73.11	65.71
system mentions × gold annotation										
CISTELL	66.64	32.92	43.03	37.30	69.99	80.93	75.06	58.28	65.93	60.78
RELAXCOR	67.60	34.53	48.28	40.26	71.70	84.41	77.54	57.49	66.41	60.14
system mentions × predicted annotation										
CISTELL	66.21	27.26	40.63	32.63	69.11	83.13	75.47	56.86	65.33	59.33
RELAXCOR	65.41	15.36	54.48	23.96	66.02	94.98	77.90	52.78	77.32	54.56
Spanish										
true mentions × gold annotation										
CISTELL	69.50	46.74	47.90	47.31	73.77	75.46	74.60	68.25	68.16	68.21
RELAXCOR	75.62	55.74	68.91	61.63	75.95	87.07	81.13	64.07	74.86	67.87
true mentions × predicted annotation										
CISTELL	68.44	40.23	42.59	41.38	72.77	75.80	74.25	64.88	66.48	65.64
RELAXCOR	74.95	58.04	65.22	61.42	76.39	83.83	79.94	64.47	71.02	67.09
system mentions × gold annotation										
CISTELL	67.99	34.84	46.26	39.74	71.10	82.12	76.22	60.70	69.14	63.63
RELAXCOR	69.72	34.86	53.30	42.15	71.40	86.65	78.29	58.32	69.92	61.57
system mentions × predicted annotation										
CISTELL	67.51	29.45	43.69	35.18	70.42	83.70	76.49	59.85	68.52	62.76
RELAXCOR	65.99	19.40	44.77	27.07	68.38	90.82	78.02	53.45	70.40	55.55
All languages										
true mentions × gold annotation										
CISTELL	69.86	45.47	46.78	46.12	74.09	76.57	75.31	65.99	68.09	66.97
RELAXCOR	76.53	56.39	68.85	62.00	77.48	87.39	82.14	65.37	74.95	68.95
true mentions × predicted annotation										
CISTELL	68.97	39.82	41.89	40.83	73.51	76.79	75.12	63.07	66.16	64.44
RELAXCOR	75.83	56.25	67.51	61.37	77.01	86.71	81.57	64.05	73.36	67.46
system mentions × gold annotation										
CISTELL	68.19	34.55	43.94	38.69	71.96	81.62	76.48	59.47	67.49	62.19
RELAXCOR	70.55	36.32	52.58	42.97	73.20	86.35	79.23	58.89	69.11	62.01
system mentions × predicted annotation										
CISTELL	67.86	29.77	41.82	34.78	71.39	83.16	76.83	58.56	66.91	61.26
RELAXCOR	67.94	20.30	51.93	29.19	69.30	92.81	79.35	54.34	74.71	57.03

Overall performances. The best system appears to be RELAXCOR, especially for English, while the measures disagree in ranking RECONCILE and CISTELL. The former is the second top system according to CEAF and B^3, whereas the latter is according to MUC and BLANC. This disagreement is associated with the opposite tendencies of the two systems: they obtain similar F_1 scores, but RECONCILE favors precision over recall, while the opposite is true for CISTELL, as the examples in the next section illustrate. The limitations of the measures in relation to the baselines become apparent again (Table 5): although RECONCILE and CISTELL only slightly outperform the B^3 and CEAF results of the SINGLETONS baseline, and generally underperform the MUC result of the ALL-IN-ONE baseline, their outputs are certainly preferable to simply classifying all the mentions as singletons, or linking them all under the same entity.

Languages. In terms of language, the best results are obtained for English, followed by Spanish and Catalan (RELAXCOR and CISTELL come close to each other if we follow the BLANC ranking). Two factors account for this difference. First, the larger number of singletons observed in English, which boosts the B^3 and CEAF performance. Second, the system that actually shows the most dramatic decrease, RELAXCOR, was originally designed with the English language in mind. As a result, it does not include language-specific features for Spanish and Catalan like whether a mention is or not an elliptical subject. The slightly worse performance in Catalan as compared with Spanish reflects the different composition summarized in Table 2 (Section 2). Despite the similarity between the two corpora, Catalan is expected to be harder given the higher distance between coreferent mentions and its lower decayed density.

Gold vs predicted, true vs system. In terms of gold versus predicted annotation, and of true versus system mentions, it emerges that the largest drop in performance is observed for the link-based measures. Performance decreases by 5–6 MUC points in the case of RELAXCOR and RECONCILE in English, but only by 2 points in the case of CISTELL, while the decrease is hardly noticeably looking at B^3. In Catalan and Spanish, the CISTELL score decreases to a larger extent than that of RELAXCOR when true mentions are used, but RELAXCOR experiences a very remarkable drop when system mentions are used. This is very likely due to the fact that RELAXCOR was not separately trained on system mentions. The system was trained for each language, but not for each evaluation scenario due to the high computational cost of learning the constraints (the only scenario tuning occurred during development).

This also explains that CISTELL comes very close to RELAXCOR for Catalan and Spanish in the *system mentions* $\times$ *gold annotation* setting, and even outperforms it in *system mentions* $\times$ *predicted annotation*. The performance decrease from true to system mentions is the expected one given the mention detection results shown in Table 6. In general, recall registers a higher decrease than precision as the true mentions that are missed cannot be recovered, while the system mentions that do not have a true counterpart can still be counterbalanced during the training stage if the system learns to classify them as singletons.

State of the art. It is not possible to compare the results of Table 7 with state-of-the-art results because different data sets were used, and because of the disagreements between the evaluation measures. All the data sets and evaluation software for this task are publicly available for anyone who wishes to assess their results, and Table 7 can be used as a baseline in the future.

6 Analysis and Discussion

The results of Table 7 are compressed into a single score for each system output, making it hard to know what are the specific strengths and weaknesses of each system, whether they perform differently for different mention types, or whether they show similar patterns. In order to reveal details that cannot be seen in the numerical table of results, and by way of an error analysis, this section breaks down the *true mentions* × *gold annotation* coreference results by mention class (Table 8), and examines specific real examples (Figures 3 and 4). This also leads us to reflect on the different evaluation measures, and consider how to choose the *best* output.

6.1 System Analysis

Breaking down the coreference score by mention class makes it easier to identify the "easy" and "hard" coreference relations, and consider whether they are system specific. To this end, we modified the four evaluation measures to obtain partial scores according to the mention classes described in Figure 2 (Section 2). Table 8 displays the results by mention class; due to lack of space, we restrict it to the *true mentions* × *gold annotation* scenario (also because in this way we avoid errors from other annotation layers) and to the scores of one mention-based measure (CEAF) and one link-based measure (BLANC).[9]

General observations about the scores by mention class. The first notable observation from Table 8 is that the best-scoring classes in English are PN_E and PN_P, that is, proper nouns with either exact or partial match. However, not all the systems behave the same in this respect: while RELAXCOR and RECONCILE perform clearly better on PN_P than on CN_E, the CISTELL scores for PN_P are lower than for CN_E and CN_P. The example in Figure 3 shows, for instance, that CISTELL is the only system that links *the Yemeni port of Aden* and *the port*. In general, RELAXCOR is a more precision-oriented system that is reluctant to corefer common nouns even if they match partially (recall also that RELAXCOR was tuned for the CEAF evaluation measure, which is a measure that favors precision more than recall as can be seen in the baselines results). In contrast, in an attempt to improve recall, CISTELL corefers more common nouns, but this results in an overall worse performance. RECONCILE

[9] Although our scores by class are similar to Stoyanov et al's (2009) MUC-RC score, a variant of MUC, we do not start from the assumption that all the coreferent mentions that do not belong to the class under analysis are resolved correctly. The results by mention class for all the scenarios and measures as well as the detailed scoring software are available at http://nlp.lsi.upc.edu/coreference/LRE-2011.

Table 8 Coreference results of the three systems broken down by mention class. CEAF and BLANC evaluation measures are reported over all the languages and in the *true mentions* × *gold annotation* setting

	English				Catalan				Spanish			
	CEAF	BLANC			CEAF	BLANC			CEAF	BLANC		
	F_1	R	P	Blanc	F_1	R	P	Blanc	F_1	R	P	Blanc
CISTELL												
PN_E	68.0	70.8	86.5	75.4	70.5	66.0	82.2	70.2	66.7	70.2	83.9	74.5
PN_P	50.0	61.6	83.6	66.1	59.6	61.2	78.2	63.9	57.0	68.0	75.0	70.7
PN_N	47.0	58.3	76.7	62.1	38.8	57.1	73.2	60.0	41.1	62.0	68.3	64.3
CN_E	64.0	62.3	87.0	66.7	65.1	68.1	78.9	71.7	67.5	71.3	80.4	74.7
CN_P	58.5	66.9	81.5	71.8	63.7	63.9	79.6	68.4	65.1	65.3	86.4	70.8
CN_N	25.6	51.2	60.7	51.2	24.8	54.4	58.8	55.4	25.3	55.2	55.9	55.5
P_1∪2	52.2	64.2	68.5	65.9	13.7	51.7	52.6	51.5	48.9	72.9	71.4	72.1
P_3G	48.1	62.2	67.8	64.1	28.9	61.1	64.2	62.4	32.5	61.6	60.7	61.1
P_3U	27.2	57.9	59.9	58.1								
P_ELL					43.8	67.6	66.0	66.8	50.0	70.5	66.6	68.1
P_REL					25.7	50.1	50.2	50.1	24.2	51.2	52.2	51.5
		R	P	F_1		R	P	F_1		R	P	F_1
SING		66.8	85.0	74.8		69.3	73.1	71.2		68.0	74.7	71.2
RELAXCOR												
PN_E	93.3	85.7	95.7	89.7	86.3	75.7	90.9	80.8	87.7	73.8	92.8	79.6
PN_P	89.8	83.1	95.8	88.2	50.6	63.9	81.6	67.5	63.6	70.9	94.7	77.4
PN_N	63.1	67.4	93.7	74.5	52.1	63.6	81.7	68.4	49.3	63.9	82.5	68.9
CN_E	64.0	66.6	93.7	72.4	70.1	66.8	87.1	71.9	70.2	68.4	87.5	73.6
CN_P	42.0	60.1	88.8	65.4	63.0	66.8	89.0	72.9	59.1	63.8	90.7	69.6
CN_N	13.7	50.4	67.6	49.6	22.3	54.8	62.3	56.3	24.7	55.1	64.6	57.1
P_1∪2	48.5	66.8	68.5	67.6	35.3	56.6	57.4	56.9	35.6	55.4	62.8	54.8
P_3G	79.1	82.8	83.9	83.3	37.5	60.4	70.0	63.1	34.3	58.7	63.9	60.4
P_3U	52.5	67.2	82.7	72.2								
P_ELL					34.2	57.3	62.3	58.6	33.9	58.5	67.2	60.4
P_REL					71.3	56.4	58.2	57.2	75.1	61.0	63.3	62.0
		R	P	F_1		R	P	F_1		R	P	F_1
SING		91.3	81.7	86.2		82.9	73.8	78.1		86.7	76.2	81.1
RECONCILE												
PN_E	83.4	75.9	93.1	81.3								
PN_P	64.8	67.2	95.4	73.9								
PN_N	35.4	55.8	82.3	59.0								
CN_E	27.3	53.8	94.3	54.0								
CN_P	6.8	50.9	92.2	50.6								
CN_N	4.6	50.2	85.1	49.2								
P_1∪2	35.1	56.8	87.4	59.3								
P_3G	19.8	53.1	85.6	52.7								
P_3U	17.7	53.0	90.6	54.2								
		R	P	F_1								
SING		96.0	69.9	80.9								

is the most precision-oriented system of the three and links a very small number of mentions (only two mentions in the example). Note that apart from PN_E and PN_P, it obtains very low scores for the other classes. This behavior could probably be changed by adjusting the value of the coreference decision threshold (set to the default 0.5 for this study) on the development set. If we rank the systems by the number of links, from highest to lowest, we obtain CISTELL > RELAXCOR > RECONCILE, but RELAXCOR seems to find the best trade-off between precision and recall.

The lowest-scoring classes for non-pronominal mentions in English are the non-exact-match ones, namely CN_N and PN_N to a lesser extent for CISTELL, and CN_N and CN_P for RELAXCOR and RECONCILE. This is to be expected as these are the mentions that require more semantic and world knowledge to be solved, and it is in accordance with previous research. The semantic features used by the three systems

are limited to NE type (i.e., whether two NE mentions belong to the same class) and WordNet (i.e., whether the two mention heads are connected by a synonymy or hypernymy relation in WordNet). In Figure 3, all the systems fail to link *the USS Cole destroyer* and *the ship*, or *a suspected terrorist attack that killed at least six sailors* and *the blast*. There seems to be a trend in that CISTELL evens out the classes of proper nouns and of common nouns, while a major strength of the other two systems is in solving proper nouns.

Languages. Although the rankings of classes in Catalan and Spanish are highly comparable with the ranking in English, they show differences that are worth mentioning. Unlike in English, RELAXCOR performs better on CN_E than PN_P in the two Romance languages. This was already the case for CISTELL in English. This might have to do with the larger percentages of CN_E but lower percentages of PN_P in Catalan and Spanish observed in Table 3. Despite the generally lower results in Catalan and Spanish, it is remarkable that the CN_N and CN_P classes obtain similar or even higher scores than English, especially for RELAXCOR. The performance drop of RELAXCOR for the Romance languages appears to be largely due to the drop in the performance for proper noun classes (as well as pronouns, discussed next).

Pronouns. In terms of pronouns, the systems behave differently: the hardest class is P_1∪2 for RELAXCOR, while it is P_3U for CISTELL and RECONCILE (but not far from P_3G for the latter). RECONCILE performs the worst for pronominal mentions. It gives again priority to precision at the expense of a very low recall. RELAXCOR stands outs especially in third-person pronouns, but the ungendered *it* pronoun poses problems for all the systems, as shown in Figure 3. In general, pronouns are harder to solve than full NPs. The scores for Catalan and Spanish are again lower than those for English, although they are not directly comparable because of the prevalence of elliptical subjects in the Romance languages. Interestingly enough, CISTELL performs better than RELAXCOR on P_ELL, while the opposite is true on P_REL. Recall that RELAXCOR did not include any language-specific feature, which probably accounts for its low performance on ellipticals. Clearly, the scores for elliptical subjects would be much lower if they were not marked as tokens in both the *gold-standard* and *predicted* annotations.

Singletons. The scores for singletons (SING) in Table 8 are computed as standard recall, precision, and F_1, because there is no need to use sophisticated coreference measures like CEAF or BLANC when we do not want to compare entities composed of more than one mention. From best to worst performance, the systems are RELAXCOR> RECONCILE> CISTELL. Again, CISTELL and RECONCILE behave the opposite in terms of recall and precision, the former showing a lower recall as it tends to link more mentions, whereas the conservative nature of RECONCILE in establishing coreference links accounts for its very large recall but very low precision for singletons. Again, the performance drop in Catalan and Spanish is larger for RELAXCOR than CISTELL. Whereas CISTELL shows no difference between the two Romance languages, RELAXCOR performs worse for Catalan.

Fig. 3 Outputs of voa_0033.txt in the *true mentions* × *gold annotation* scenario. Coreferent mentions share the same subscript number. The colors in the GOLD identify coreferent mentions, while the colors in the system outputs identify the coreferent mentions that are correctly (in blue) and wrongly (in red) output by the system

GOLD

[US officials]$_1$ have launched [an investigation into [a suspected terrorist attack that killed [at least six sailors] on [a US warship in [Yemen]]$_3$]$_2$]. [They]$_1$ say [a suicide boat]$_4$ exploded near [the USS Cole destroyer]$_3$ as [it]$_3$ refueled [early Thursday] in [the Yemeni port of [Aden]]$_5$. [The blast]$_2$ tore [a huge gap in [[the ship's]$_3$ side]]. [At least 10 other crewmembers] are missing more than 30 were injured. [Defense Secretary William Cohen] says if [a formal investigation] shows [it]$_2$ was [a terrorist attack]. [Dispatch investigators to learn what was responsible for [the blast]]$_2$. [Tragedies that test [[[our]$_6$ strength] and [[our]$_6$ resolve]] occur and have occurred in [the past]]. [We]$_6$ have passed [[the test]$_7$ and measuring up to [those responsibilities]] and [we]$_6$ will pass [that test]$_7$ once again [today]. [Chief of [Naval Operations] Admiral Vern Clark] says [the bombers] were aboard [a small boat that was assisting in [the refueling of [the destroyer]$_3$ at [the port]$_5$]]$_4$.

CISTELL

83.92% CEAF	59.25% MUC	88.46% B^3	51.16% BLANC

[US officials]$_1$ have launched *[an investigation into [a suspected terrorist attack that killed at least six sailors on a US warship in Yemen]$_2$]$_8$*. [They]$_1$ say a suicide boat exploded near [the USS Cole destroyer]$_3$ as *[it]$_9$* refueled early Thursday in [the Yemeni port of Aden]$_5$. [The blast]$_{10}$ tore a huge gap in [the ship's]$_9$ side. At least 10 other crewmembers are missing more than 30 were injured. Defense Secretary William Cohen says if *[a formal investigation]$_8$* shows [it]$_{10}$ was *[a terrorist attack]$_2$*. Dispatch investigators to learn what was responsible for the blast. Tragedies that test [our]$_6$ strength and [our]$_6$ resolve occur and have occurred in the past. [We]$_6$ have passed [the test]$_7$ and measuring up to those responsibilities and [we]$_6$ will pass [that test]$_7$ once again today. Chief of Naval Operations Admiral Vern Clark says the bombers were aboard a small boat that was assisting in the refueling of [the destroyer]$_3$ at [the port]$_5$.

RECONCILE

75.00% CEAF	12.50% MUC	85.12% B^3	6.89% BLANC

US officials have launched an investigation into a suspected terrorist attack that killed at least six sailors on a US warship in Yemen. They say a suicide boat exploded near the USS Cole destroyer as it refueled early Thursday in the Yemeni port of Aden. The blast tore a huge gap in the ship's side. At least 10 other crew members are missing more than 30 were injured. Defense Secretary William Cohen says if a formal investigation shows it was a terrorist attack. Dispatch investigators to learn what was responsible for the blast. Tragedies that test [our]$_6$ strength and [our]$_6$ resolve occur and have occurred in the past. We have passed the test and measuring up to those responsibilities and we will pass that test once again today. Chief of Naval Operations Admiral Vern Clark says the bombers were aboard a small boat that was assisting in the refueling of the destroyer at the port.

RELAXCOR

83.92% CEAF	54.54% MUC	89.39% B^3	46.15% BLANC

[US officials]$_1$ have launched an investigation into a suspected terrorist attack that killed at least six sailors on a US warship in Yemen. [They]$_1$ say a suicide boat exploded near the USS Cole destroyer as *[it]$_2$* refueled early Thursday in the Yemeni port of Aden. [The blast]$_2$ tore a huge gap in the ship's side. At least 10 other crewmembers are missing more than 30 were injured. Defense Secretary William Cohen says if a formal investigation shows it was a terrorist attack. Dispatch investigators to learn what was responsible for [the blast]$_2$. Tragedies that test [our]$_6$ strength and [our]$_6$ resolve occur and have occurred in the past. [We]$_6$ have passed [the test]$_7$ and measuring up to those responsibilities and [we]$_6$ will pass [that test]$_7$ once again today. Chief of Naval Operations Admiral Vern Clark says the bombers were aboard a small boat that was assisting in the refueling of the destroyer at the port.

Fig. 4 Outputs of wsj_1245.txt in the *true mentions* × *gold annotation* and *system mentions* × *gold annotation* scenarios (RECONCILE is not shown as it only outputs singletons). Coreferent mentions share the same subscript number. The colors in the GOLD identify coreferent mentions, while the colors in the system outputs identify the coreferent mentions that are correctly (in blue) and wrongly (in red) output by the system

GOLD
[Consumers Power [Co.]$_2$]$_1$ filed with **[the Michigan Public Service Commission]$_2$** [a contract to buy [power] from **[the Palisades nuclear plant]$_3$** under [a proposed new ownership arrangement for **[the plant]$_3$**]]. [**[Consumers Power]$_1$** and [Bechtel Power [Corp.]]] [last year] announced [a joint venture to buy **[the plant, currently owned completely by [the utility]$_2$]$_3$**].

CISTELL *true mentions* × *gold annotation*			
75.00% CEAF	44.44% MUC	84.25% B^3	46.15% BLANC
[Consumers Power [Co.]]$_1$ filed with [the Michigan Public Service Commission] [a contract to buy [power] from **[the Palisades nuclear plant]$_2$** under [a proposed new ownership arrangement for **[the plant]$_2$**]]. **[Consumers Power and** *[Bechtel Power [Corp.]]$_1$*]$_1$ [last year] announced [a joint venture to buy **[the plant, currently owned completely by [the utility]]$_2$**].			

CISTELL *system mentions* × *gold annotation*			
75.00% CEAF	50.00% MUC	87.05% B^3	54.54% BLANC
Consumers Power [Co.] filed with [the Michigan Public Service Commission] [a contract to buy *[power]$_1$* from **[the [Palisades] nuclear plant]$_2$** under [a proposed new ownership arrangement for **[the plant]$_2$**]]. *[Consumers Power and Bechtel Power [Corp.]]$_1$* [last year] announced [a joint venture to buy **[the plant, currently owned completely by [the utility]]$_2$**].			

RELAXCOR *true mentions* × *gold annotation*			
81.25% CEAF	57.14% MUC	88.37% B^3	44.44% BLANC
[Consumers Power [Co.]]$_1$ filed with [the Michigan Public Service Commission] [a contract to buy [power] from **[the Palisades nuclear plant]$_2$** under [a proposed new ownership arrangement for [the plant]]]. [**[Consumers Power]$_1$** and [Bechtel Power [Corp.]]] [last year] announced [a joint venture to buy **[the plant, currently owned completely by [the utility]]$_2$**].			

RELAXCOR *system mentions* × *gold annotation*			
81.25% CEAF	57.14% MUC	89.65% B^3	60.00% BLANC
Consumers Power [Co.] filed with [the Michigan Public Service Commission] [a contract to buy [power] from **[the [Palisades] nuclear plant]$_1$** under [a proposed new ownership arrangement for **[the plant]$_1$**]]. [Consumers Power and Bechtel Power [Corp.]] [last year] announced [a joint venture to buy **[the plant, currently owned completely by [the utility]]$_1$**].			

6.2 Measure Analysis

As it was the case with the results in Table 7, Table 8 also reveals various contradictions between the evaluation measures in scoring the different outputs. CISTELL, for instance, obtains a larger score for CN_E than CN_P according to CEAF but smaller according to BLANC in English. The same tendency occurs with RELAXCOR in Catalan, but to a lesser extent. In contrast, CEAF always shows that the CN_E class is easier than the CN_P class. It is not straightforward to explain the reason for this. It could be due to one of the drawbacks of CEAF: given that it establishes the best one-to-one entity alignment, if a CN_P mention is correctly linked to a preceding mention but this does not fall under the "best one-to-one alignment," then CEAF does not reward this link correctly solved at a local level. The examples in Figures 3 and 4 also show disagreements between the scores. In the first example, CISTELL and RELAXCOR obtain the same CEAF score in both cases, whereas B^3 ranks RELAXCOR first, and MUC and BLANC rank CISTELL first. The link-based measures put more

emphasis on correct coreference links (even if it is at the expense of incorrect ones), whereas the score of mention-based measures decreases rapidly in the presence of incorrect coreference links, as singletons count as an entity per se.

The example in Figure 4, where we can compare the outputs using true and system mentions in the gold scenario, also reveals the different sensitivities of each measure. In this example, singletons are marked within square brackets to better illustrate the true-system versus system-mention outputs. Surprisingly, the CEAF score stays the same in the two scenarios for CISTELL and RELAXCOR, also the MUC score for the latter system, while the rest of measures rank better the output in *system mentions* $\times$ *gold annotation*. This is a very short document and (in)correctly solving a single link can make a big difference. In the case of CISTELL, for instance, although the first output might seem better at first sight, it links wrongly the mention *Bechtel Power Corp.* together with *Consumers Power Co.* and *Consumers Power and Bechtel Power Corp*. The second output also links wrongly one mention, *power*, but only with another mention (*Consumers Power and Bechtel Power Corp.*). A similar issue happens in the RELAXCOR outputs. Notice again the more precision-oriented nature of RELAXCOR versus the more recall-oriented nature of CISTELL.

Both examples illustrate the difficulty of evaluating coreference resolution: Do we prefer few but good links rather than more recall but less precision? The different results reported in this study indicate that the extreme cases are clear and the different measures agree, but in-between cases are not so clear and this is where the measures (and human annotators) often disagree. There is no correct answer in absolute terms, but a possible range of answers, and evaluation is very task-specific. Depending on the intended application, one or another answer will be preferred. If recall matters more than precision, it is wiser to use a link-based measure, whereas if precision matters more than recall, then it is wiser to use a mention-based measure. Although the tradition in coreference resolution evaluation has been to use intrinsic measures, the coreference community should start applying more extrinsic evaluation methodologies.

7 Conclusions

This paper has presented a multi-dimensional empirical study of coreference resolution. The analyzed dimensions include:

- *Multilinguality*: by using English, Catalan and Spanish corpora.
- *Approaches to coreference resolution*: by including the entity-mention system CISTELL, and the mention-pair models RELAXCOR and RECONCILE.
- *Evaluation measures*: by evaluating with the mention-based measures B^3 and CEAF, and the link-based MUC and BLANC.
- *Evaluation scenarios*: by training with *gold* versus *predicted* input information, and with *true* versus *system* mentions.

Departing from the definition and materials of the SemEval-2010 Task 1 (Recasens et al, 2010), this study slightly reduces the complexity (less languages and systems, less evaluation settings), but produces a complete study, fixing also some

of the design errors of the SemEval-2010 task and including a detailed discussion of system outputs and examples.

In Section 2, a first study of the corpora is conducted, presenting statistics on the coreference annotation. The statistics reveal remarkable differences between the English and Romance corpora with respect to the proportion of singletons, the density of entity mentions, and the distribution of mentions by classes. Some of these phenomena have a linguistic interpretation, and they are later shown to influence the performance of the coreference resolution systems in each language.

Section 4 presents the full table with the results of all the systems across languages, settings and measures. This table represents the completion of the main analysis from the SemEval-2010 task. In general, systems perform better for English, followed by Spanish and Catalan. Reasons for this include properties of the corpora (e.g., the proportion of singletons) as well as the original language for which a system was originally developed.

All the evaluation measures agree that RELAXCOR performs consistently better, especially for English. However, when comparing CISTELL and RECONCILE, they disagree in the overall results. Since the two systems have very different behavior in terms of precision and recall, this reveals that the measures tend to reward different aspects of quality. Therefore, unless the differences between systems are large, a single evaluation measure is not enough to allow for general quality comparisons. It seems necessary to apply more than one evaluation measure to make an informed decision about which measure fits best in each situation. Going one step further, the authors believe that the most reasonable way to go would be to use task-specific measures, both for developing and comparing systems, in order to provide valuable insights into how coreference resolution impacts the performance of NLP applications.

This work also highlights the limitation of the measures in relation to the extreme baseline systems. We show that according to some measures, there are cases in which real systems perform comparably or even worse than simply considering all the mentions as singletons (SINGLETONS), or joining them all into a single entity (ALL-IN-ONE). Another subtle aspect of the difficulty of evaluating coreference is the mapping between true and system mentions when they do not coincide. As shown in this work, the adjustment proposed by Cai and Strube (2010) can be generalized to all the measures in order to avoid unfair evaluations. However, this adjustment turned out to have the counterintuitive effect of making the ALL-IN-ONE baseline based on system mentions obtain higher scores than the one based on true mentions.

Moreover, it was hard to draw reliable conclusions regarding the comparison between *mention-pair* (RELAXCOR and RECONCILE) and *entity-mention* (CISTELL) approaches to coreference resolution. At first sight, it seems that mention-pair systems are biased toward high precision, while entity-mention systems are biased toward high recall. However, with only three systems it is not possible to determine whether performance differences are attributable to the intrinsic properties of each approach or to aspects from other dimensions. On the other hand, it seems that the factor that matters most for adapting a system to a specific corpus or language is adjusting the right parameters, learning features, and training conditions, rather than

the approach or architecture itself. At this stage, we leave these two issues for further research.

In Section 6, a more detailed analysis is performed by evaluating the systems according to mention classes. Previous research has emphasized that head matching mentions are the easiest to resolve, and we add to this observation that mentions headed by proper nouns are easier to resolve than mentions headed by common nouns. Obviously, coreferent mentions with different heads remain a major challenge. Some examples are also provided to back up some of the quantitative observations from the previous section with regard to the trends for each system and evaluation measure.

Last but not least, an additional valuable contribution of this work is the collection of resources that it has made available to the community, including the updated versions of the corpora and the scoring software, and the system outputs in a user-friendly format for viewing. We hope that these resources will be of interest to the coreference community, and that they will become benchmarks for future evaluations.

Acknowledgements This work was partially funded 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). It also received financial support from the Seventh Framework Programme of the EU (FP7/2007- 2013) under GAs 247762 (FAUST) and 247914 (MOLTO), and from Generalitat de Catalunya through a Batista i Roca project (2010 PBR 00039). We are grateful to the two anonymous reviewers of this paper. Their insightful and careful comments allowed us to significantly improve the quality of the final version of this manuscript.

References

Abad A, Bentivogli L, Dagan I, Giampiccolo D, Mirkin S, Pianta E, Stern A (2010) A resource for investigating the impact of anaphora and coreference on inference. In: Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp 128–135

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

Bagga A, Baldwin B (1998) Algorithms for scoring coreference chains. In: Proceedings of the Linguistic Coreference Workshop at LREC 98, Granada, Spain, pp 563–566

Bengtson E, Roth D (2008) Understanding the value of features for coreference resolution. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, USA, pp 294–303

Cai J, Strube M (2010) Evaluation metrics for end-to-end coreference resolution systems. In: Proceedings of the annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2010), Tokyo, Japan, pp 28–36

Chambers N, Jurafsky D (2008) Unsupervised learning of narrative event chains. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-HLT 2008), Columbus, USA, pp 789–797

Civit M, Martí MA (2005) Building Cast3LB: A Spanish Treebank. Research on Language and Computation 2(4):549–574

Daelemans W, Buchholz S, Veenstra J (1999) Memory-based shallow parsing. In: Proceedings of the Conference on Natural Language Learning (CoNLL 1999), Bergen, Norway, pp 53–60

Daumé H, Marcu D (2005) A large-scale exploration of effective global features for a joint entity detection and tracking model. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), Vancouver, Canada, pp 97–104

Denis P, 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, Weischedel R (2004) The Automatic Content Extraction (ACE) program – Tasks, data, and evaluation. In: Proceedings of the 4th Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp 837–840

Finkel J, Manning C (2008) Enforcing transitivity in coreference resolution. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-HLT 2008), Columbus, USA, pp 45–48

Gerber M, Chai JY (2010) Beyond NomBank: A study of implicit arguments for nominal predicates. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, pp 1583–1592

Heim I (1983) File change semantics and the familiarity theory of definiteness. In: BŁuerle R, Schwarze C, von Stechow A (eds) Meaning, Use, and Interpretation of Language, Mouton de Gruyter, Berlin, Germany, pp 164–189

Hirschman L, Chinchor N (1997) MUC-7 Coreference Task Definition – Version 3.0. In: Proceedings of the 7th Message Understanding Conference (MUC-7), Fairfax, USA

Hummel RA, Zucker SW (1987) On the foundations of relaxation labeling processes. In: Fischler MA, Firschein O (eds) Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, Morgan Kaufmann Publishers Inc., San Francisco, USA, pp 585–605

Lundquist L (2007) Lexical anaphors in Danish and French. In: Schwarz-Friesel M, Consten M, Knees M (eds) Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference, John Benjamins, Amsterdam, Netherlands, pp 37–48

Luo X (2005) On coreference resolution performance metrics. In: Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005, Vancouver, Canada, pp 25–32

Luo X, Ittycheriah A, Jing H, Kambhatla N, Roukos S (2004) A mention-synchronous coreference resolution algorithm based on the Bell tree. In: Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, pp 21–26

McCarthy JF, Lehnert WG (1995) Using decision trees for coreference resolution. In: Proceedings of The 1995 International Joint Conference on AI (IJCAI 1995), Montreal, Canada, pp 1050–1055

Mirkin S, Berant J, Dagan I, Shnarch E (2010) Recognising entailment within discourse. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, pp 770–778

Morton TS (1999) Using coreference in question answering. In: Proceedings of the 8th Text REtrieval Conference (TREC-8), pp 85–89

Ng V (2010) Supervised noun phrase coreference research: the first fifteen years. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, pp 1396–1411

Ng V, Cardie C (2002) Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, USA, pp 104–111

Nicolov N, Salvetti F, Ivanova S (2008) Sentiment analysis: Does coreference matter? In: Proceedings of the Symposium on Affective Language in Human and Machine, Aberdeen, UK, pp 37–40

Orasan C, Cristea D, Mitkov R, Branco A (2008) Anaphora Resolution Exercise: An overview. In: Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp 28–30

Padró L (1998) A hybrid environment for syntax–semantic tagging. PhD thesis, Dep. Llenguatges i Sistemes Informaics. Universitat Politècnica de Catalunya, Barcelona, Spain

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, 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, Los Angeles, USA, pp 87–95

Popescu A, Etzioni O (2005) Extracting product features and opinions from reviews. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), Vancouver, Canada, pp 339–346

Popescu-Belis A, Robba I, Sabah G (1998) Reference resolution beyond coreference: a conceptual frame and its application. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics joint with the International Conference on Computational Linguistics (COLING-ACL 1998), Montreal, Canada, pp 1046–1052

Pradhan S, Hovy E, Marcus M, Palmer M, Ramshaw L, Weischedel R (2007) OntoNotes: A unified relational semantic representation. In: Proceedings of the International Conference on Semantic Computing (ICSC 2007), Irvine, USA, pp 517–526

Pradhan S, Ramshaw L, Marcus M, Palmer M, Weischedel R, Xue N (2011) CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes. In: Proceedings of the Conference on Natural Language Learning (CoNLL 2011): Shared Task, Portland, USA, pp 1–27

Quinlan J (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, Massachusetts, USA

Rahman A, Ng V (2009) Supervised models for coreference resolution. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), Suntec, Singapore, pp 968–977

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

Recasens M (2010) Coreference: Theory, Annotation, Resolution and Evaluation. PhD thesis, University of Barcelona, Barcelona, Spain

Recasens M, Hovy E (2009) A deeper look into features for coreference resolution. In: Devi SL, Branco A, Mitkov R (eds) Anaphora Processing and Applications (DAARC 2009), Springer-Verlag, Berlin, Germany, LNAI, vol 5847, pp 29–42

Recasens M, Hovy E (2011) BLANC: Implementing the Rand Index for coreference evaluation. Natural Language Engineering 17(4):485–510

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

Recasens M, Màrquez L, Sapena E, Martí MA, Taulé M, Hoste V, Poesio M, Versley Y (2010) Semeval-2010 task 1: Coreference resolution in multiple languages. In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010), Uppsala, Sweden, pp 1–8

Ruppenhofer J, Sporleder C, 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), Uppsala, Sweden, pp 45–50

Sapena E, Padró L, Turmo J (2010a) A global relaxation labeling approach to coreference resolution. In: Proceedings of 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, pp 1086–1094

Sapena E, Padró L, Turmo J (2010b) Relaxcor: A global relaxation labeling approach to coreference resolution. In: Proceedings of the ACL Workshop on Semantic Evaluations (SemEval-2010), Uppsala, Sweden, pp 88–91

Soon WM, Ng HT, Lim DCY (2001) A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4):521–544

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

Stoyanov V, Gilbert N, Cardie C, Riloff E (2009) Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009), Suntec, Singapore, pp 656–664

Stoyanov V, Cardie C, Gilbert N, Riloff E, Buttler D, Hysom D (2010) Coreference resolution with Reconcile. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, pp 156–161

Versley Y, Ponzetto S, Poesio M, Eidelman V, Jern A, Smith J, Yang X, Moschitti A (2008) BART: A modular toolkit for coreference resolution. In: Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp 962–965

Vicedo JL, Ferrández A (2006) Coreference in Q&A. In: Strzalkowski T, Harabagiu S (eds) Advances in Open Domain Question Answering, Text, Speech and Language Technology, vol 32, Springer-Verlag, Berlin, Germany, pp 71–96

Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L (1995) A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6), pp 45–52

Wick M, Culotta A, Rohanimanesh K, McCallum A (2009) An entity based model for coreference resolution. In: Proceedings of The SIAM Data Mining Conference (SDM 2009), Reno, USA, pp 365–376