Annotating Near-Identity from Coreference Disagreements

Marta Recasens*, **M. Antònia Martí***, **Constantin Orasan**†

*CLiC - University of Barcelona
Gran Via 585, Barcelona, Spain
{mrecasens, amarti}@ub.edu

†University of Wolverhampton
Stafford Street, Wolverhampton, UK
c.orasan@wlv.ac.uk

Abstract

We present an extension of the coreference annotation in the English NP4E and the Catalan AnCora-CA corpora with near-identity relations, which are borderline cases of coreference. The annotated subcorpora have 50K tokens each. Near-identity relations, as presented by Recasens et al. (2010; 2011), build upon the idea that identity is a continuum rather than an either/or relation, thus introducing a middle ground category to explain currently problematic cases. The first annotation effort that we describe shows that it is not possible to annotate near-identity explicitly because subjects are not fully aware of it. Therefore, our second annotation effort used an indirect method, and arrived at near-identity annotations by inference from the disagreements between five annotators who had only a two-alternative choice between coreference and non-coreference. The results show that whereas as little as 2–6% of the relations were explicitly annotated as near-identity in the former effort, up to 12–16% of the relations turned out to be near-identical following the indirect method of the latter effort.

Keywords: coreference, near-identity, corpus annotation

1. Introduction

A number of coreferentially annotated corpora have become available since the mid 1990s: MUC (Hirschman and Chinchor, 1997), ACE (Doddington et al., 2004), GNOME (Poesio, 2004), NP4E (Hasler et al., 2006), AnCora-CA (Recasens and Martí, 2010), TüBa-D/Z (Hinrichs et al., 2005), etc. While they differ in the range of mentions that are annotated and other respects, they all share the understanding of coreference as an either/or relation: two mentions *either* refer to the 'same' entity *or* do not. It is not controversial that *Zukhurov* and *he* refer to the same person in (1), but annotators are divided as to whether a case like (2) is coreferent: *panic-stricken residents carrying a few belongings on their heads* and *people in panic* refer to the same group, but they are just 'almost' the same in terms of members, as the two sets might not include the exact same members.

(1) **Zukhurov** was seized when **he** went to the rebels' base camp on Friday.

(2) Aid workers said **panic-stricken residents carrying a few belongings on their heads** fled on Tuesday for safer areas outside the city in the wake of at least 500 who left on Monday. "**People in panic** are fleeing in all directions," another aid worker said.

Interestingly enough, the ACE and OntoNotes corpora (Doddington et al., 2004; Pradhan et al., 2007) annotate the same texts completely independent of each other. Comparison of the annotation shows a significant number of disagreements.

After laying out the theoretical foundations of near-identity[1] in Recasens et al. (2010; 2011), a recent experiment that we conducted online (Recasens et al., in prep) has provided psychological evidence of this middle-ground category of near-identity (at least for English and Catalan). When subjects were asked to classify two highlighted NPs as either identity or non-identity, they tended to disagree on the same NP pairs. When participants were asked to rate the highlighted NPs from 1 to 4, clear cases of coreference or non-coreference were systematically rated as 4 and 1, respectively, whereas borderline identity relations received responses over the entire range. The annotation presented here is a follow-up to this work, and it is also supported by the findings of Poesio and Artstein (2008), Versley (2008), and van Deemter (2010). We develop an English and a Catalan 50K-word corpus annotated with near-identity relations. We take advantage of two existing coreferentially annotated corpora, namely the English NP4E (Hasler et al., 2006) and the Catalan AnCora-CA (Recasens and Martí, 2010).

In this paper, we present two annotation efforts. The first effort that we describe shows that it is not possible to annotate near-identity explicitly because subjects are not fully aware of it. Therefore, our second effort used an indirect method, and arrived at near-identity annotations by inference from the disagreements between five annotators who had only a two-alternative choice between coreference and non-coreference. A similar approach was taken by Poe-

[1] The term *near-identity* refers to the identity relationship holding between two referents. By extension, and for simplicity, we also use it to refer to the relationship between two NPs whose referents are near-identical, with the understanding that it is supposed to mean *near-coreference*.

sio and Artstein (2008), who found that annotators often missed the ambiguity of a mention, but that it was possible to identify ambiguity implicitly when sets of annotators chose different antecedents for a single mention. Our results show that whereas as little as 2–6% of the relations were explicitly annotated as near-identity in the former effort, up to 12–16% of the relations turned out to be near-identical following the indirect method in the latter effort.

The task of coreference resolution has attracted considerable interest in recent years, and substantial progress has been made (Lee et al., 2011; Haghighi and Klein, 2010). However, it still remains a largely unsolved problem requiring further work, as it "does not seem to be following the same kind of learning curve that we are used to with other problems of this sort" (Pradhan et al., 2011). Near-identity cases, which are borderline cases of coreference, introduce noise and adversely affect the performance of automatic systems. The more comprehensive resources resulting from our work can be used to train coreference resolution systems that are able to discriminate clear identity of reference from near-identity. A system that focuses on *clear* identity can learn a more accurate—and higher performing—coreference resolution model.

In the rest of this paper, we present background and related work (Section 2), describe briefly the NP4E and AnCora-CA corpora (Section 3), and compare the results of the tasks of annotating near-identity both explicitly and implicitly (Sections 4 and 5, respectively). Section 6 concludes by drawing general conclusions.

2. Background

It is well known that coreference annotation tasks are problematic (van Deemter and Kibble, 2000; Versley, 2008) and many issues remain still unsolved. As pointed out by Zaenen (2006), coreference phenomena are much less well understood than constituent structure. The practical approach to circumvent lengthy theoretical considerations and nuances has been to *simplify* the task by telling annotators that "two NPs are coreferent if they refer to the same entity," as defined in the MUC annotation guidelines (Hirschman and Chinchor, 1997). The ACE and OntoNotes annotation guidelines get a little bit more specific and comment on cases such as metonymy, but many other open questions are left undiscussed. High inter-annotator agreement can still be achieved as a result of weeks of training, but the final product is just *hiding* the intricacies behind coreference. Evidence for this comes from comparing the texts from the TDT collection that are shared by the ACE and OntoNotes corpora. Each corpus separately claims high inter-annotator agreement, yet (3-a) shows the annotations of ACE, and these do not coincide with the annotations of OntoNotes shown in (3-b).

(3) a. On homecoming night **Postville** feels like Hometown, USA, but a look around this town of 2,000 shows **it**'s become a miniature Ellis Island [...] For those who prefer the old Postville, Mayor John Hyman has a simple answer.

b. On homecoming night **Postville** feels like Hometown, USA, but a look around this town of 2,000 shows **it**'s become a miniature Ellis Island [...] For those who prefer **the old Postville**, Mayor John Hyman has a simple answer.

Hasler and Orasan (2009) find similar problems when annotating coreference between events. One of the many examples they discuss involves *the towns* and *the rebels*: although one would say that they are not coreferential, they both are targets of the attack event by being in the same place. This kind of indirect referential relations are in line with Poesio and Artstein (2008)'s observation that mentions form more complex structures than equivalence sets indicating identity of reference. Poesio and Artstein (2008) allow multiple antecedents for ambiguous anaphoric expressions, but do not mention how to annotate those cases where the two referents are *almost*—but not *fully*—the same.

Our idea of separating the *clear* cases of coreference from the near-identity ones—which are a smaller number, but still important—aims to delimit the coreference relations that cause most annotation disagreements. In other areas like textual entailment and machine translation, a distinction between *clear* and *unsure* cases of text alignment has also been followed. In the Microsoft Research alignment annotation of the RTE 2006 corpus (Brockett, 2007), every alignment link is marked as "sure" or "possible." This makes it possible to evaluate automatic systems more accurately by measuring precision against both sure and possible links, but measuring recall against only sure links.

The coreference annotation effort that we present tries to overcome the annotation inconsistencies entailed by the assumption that coreference is an either/or relation. Our theoretical approach, presented in Recasens et al. (2011), is based on three main ideas: (i) coreference phenomena occur between discourse entities rather than real-world entities, (ii) these entities are constructed by language speakers, hence they can be individuated at different granularity levels, and (iii) pragmatic factors play a key role in helping both speaker and reader set the appropriate granularity level for each discourse entity.

3. The NP4E and AnCora-CA Corpora

The NP4E corpus (Hasler et al., 2006) contains approximately 50,000 tokens from the Reuters corpus (Rose et al., 2002) fully annotated with NP coreference and partially annotated with event coreference. It was the result of a project whose goal was to develop a set of annotation guidelines for NP and event coreference for newswire texts in the domain of terrorism and security. The documents were selected according to five topics (i.e., Bukavu bombing, Peru hostages, Tajikistan hostages, Israel suicide bomb, China-Taiwan hijack). NP4E was analyzed using the Conexor's parser (Tapanainen and Järvinen, 1997), but only tokenization was used for the annotation process. Markables corresponded with NPs, and were identified manually at all the levels of embedding, and including all the modifiers of an NP in the markable.[2]

[2] http://clg.wlv.ac.uk/projects/NP4E/

The AnCora-CA corpus[3] (Recasens and Martí, 2010) contains 400,000 words annotated with coreference information on top of manually annotated grammatical relations, argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses (Taulé et al., 2008). AnCora-CA comprises newspaper and newswire articles from *El Periódico* newspaper, and the ACN news agency. Markables were identified according to the already existing syntactic annotations. All the NPs were considered to be markables but, unlike NP4E, markables excluded non-referring NPs such as appositive phrases, nominal predicates, negated NPs, and NPs within idioms. Also, unlike NP4E, relative pronouns were included as markables.

3.1. Entities

To make the markup scheme uniform between the two corpora and facilitate the annotators' job, we adopted the annotation format of AnCora-CA. Coreferent mentions are identified with an `entity` attribute, and mentions that refer to the same entity receive the same entity ID value (4).

(4) [**Zukhurov**]$_{\text{entity=“entity2”}}$ was seized when [**he**]$_{\text{entity=“entity2”}}$ went to the rebels' base camp on Friday.

3.2. Coreference relations

The second and subsequent mentions in a coreference chain include a `coreftype` attribute that specifies the type of relation with the previous mention. Although NP4E and AnCora-CA identify different types of relation, only the "ident" (i.e., identity) value was relevant for the annotation of near-identity. Thus, the NP4E values of "ident", "synonym", "generalisation" and "specialisation" all became "ident" for consistency purposes (5).

(5) [**Peruvian President Alberto Fujimori, almost two months into a stand-off with Marxist rebels**]$_{\text{entity=“entity5”}}$, said on Sunday [**he**]$_{\text{entity=“entity5” coreftype=“ident”}}$ remained calm.

AnCora-CA further specifies "pred" (i.e., predicative) and "dx" (i.e., discourse deixis) relations, but these were not eligible for near-identity annotation. For consistency purposes between the two corpora, the predicative relations that had been annotated as "ident" in NP4E were replaced with "pred" during the annotation process that we present in the next section.

4. Explicit Near-Identity Annotation

Due to the time-consuming nature of annotation tasks, our first idea for annotating near-identity was to take advantage of the existing coreference tags in NP4E and AnCora-CA, and add near-identity tags on top of them.

Two linguists with previous experience on annotating coreference were chosen for the annotation, one for each language. The whole NP4E corpus was annotated, and we randomly selected the same number of tokens (i.e., 50K) from AnCora-CA for annotation. Table 1 shows the number of tokens, documents, and coreference relations in the annotated data sets. The larger number of coreference relations in the English corpus is very likely due to the fact that the English documents are longer which also means they are more likely to contain more coreference relations.

[3] http://clic.ub.edu/corpus/en/

	NP4E	AnCora-CA
Tokens	49,279	51,622
Documents	94	176
Coreference relations	6,285	5,510

Table 1: Number of tokens, documents, and coreference relations in the data sets from NP4E and AnCora-CA.

4.1. Annotation guidelines

Annotators were asked to review all the mentions annotated as coreferent, and assign the `identdegree` attribute according to the identity degree between the referent of the mention under analysis and the referent of the immediately preceding mention in the same entity. Given that it could be the case that there was near-identity between the referents of two NPs that had not been initially annotated as coreferent (precisely due to their borderline nature), annotators were also asked to review these NPs. The three possible values of the `identdegree` attribute are:

- 3 (TOTAL IDENTITY)
  Both mentions refer to exactly the same referent.

  (6) Vincent told **the Toronto Star** that Canada and many other countries would be interested in finding a foreign destination for the Marxist rebels. "The Peruvians are obviously looking around," he told **the newspaper**.

  Nothing in the text of (6) indicates any difference between the two mentions, so there is total identity.

- 2 (STRONG NEAR-IDENTITY)
  Higher identity degree. The two referents are almost the same, they have much in common, even though they are not identical from a strict perspective. Although the two referents are likely to present differences, they tend to be treated as identical for practical communication.

  (7) Fujimori told the news conference that Peruvian police would not continue **their campaign of provoking the rebels**. [...] He said **the police action on Tuesday, which drew MRTA fire**, was untimely and inappropriate.

  The first mention seems to refer to an entity that continues in the future, including several instances of provoking the rebels. However, since the text is saying that this campaign will not be continued, it can be interpreted as just referring to the police action that occurred on Tuesday. Technically though, they refer to slightly different entities.

- 1 (WEAK NEAR-IDENTITY)
  Lower identity degree. The two referents are very

similar, but there is at least one feature (see the `nearidentfeature` values below) that clashes. On this ground, they are not interpreted as coreferent.

(8) Hutu hardliners in refugee camps in **Zaire** plotted revenge attacks on the Tutsi-led government that later took power in Rwanda, stoking tensions in **the region**.

In this case, the first mention refers to the parts of Zaire where there are refugee camps of Hutu hardliners, whereas the second mention refers to a wider physical location: the different areas where tension is present, which probably include the first and others.

Additionally, when an identity degree of either "1" or "2" was assigned, annotators were asked to specify the reason for the near-identity degree. The `nearidentfeature` attribute can take six values depending on the feature or dimension in which the second mention differs from the first. These features were inspired by our previous work (Recasens et al., 2011) and we thought that they would help annotators recognize cases of near-identity. If the two referents differ in more than one respect, only the most relevant mark is annotated.

- ROLE. A specific role played by a person is distinguished from his/her other facets.

  (9) "Your father was the greatest" commented an anonymous old lady while she was shaking Alessandro's hand —Gassman's best known son. "I will miss **the actor**, but I will be lacking **my father** especially," he said.

- REPRESENTATION. One mention is the representation of the other one in the format of a picture, book, movie, impression, etc.

  (10) Aznar reported that **the joint statement that he made with the British Prime Minister, Tony Blair,** has been sent to the acting president of the UE. The socialist spokesman of Occupation, José Antonio Griñán, described **the document** as "a useless note."

- GPE (geopolitical entity). The two mentions refer to different facets of the same geopolitical entity, but it is not possible to annotate them separately as the distinction becomes blurry throughout the text.

  (11) The government accuses the rebels of waging a proxy war for **Rwanda**, Burundi and Uganda, which deny involvement. ... Any ceasefire with the regular armies of Uganda, **Rwanda** and Burundi must be linked to the withdrawal of all foreign troops from Zaire.

- COMPONENTS. The two referents are not composed of the exact same members or parts, although they are used as if they were the 'same' entity. The boundaries of the entity are often not clear. This is a very frequent case for coreference relations between plural NPs.

  (12) Hutu hardliners in refugee camps in **Zaire** plotted revenge attacks on the Tutsi-led government that later took power in Rwanda, stoking tensions in **the region**.

- LOCATION. It is the same entity, but instantiated in two different locations.

  (13) In Lleida **the prices** increased 0.1%, and in Girona **they** remained stable.

- TIME. Similar to the previous case, but for time. The same entity is instantiated in two different temporal contexts.

  (14) We are still at a preliminary stage in **the conversations**. ... Government negotiator Domingo Palermo laid out the official position, but no agenda was agreed for **the upcoming talks**.

- OTHER. Any feature different from the above.

  (15) Afghanistan's former military chief Ahmad Shah Masood will try on Monday to mediate **an end to the week-old hostage crisis in the former Sovier republic of Tajikistan**. ... Predictions of **a quick end to the crisis** have been dashed in the past.

4.2. Results

Table 2 shows the distribution of `identdegree` tags that our annotators incorporated. To our surprise, the percentage of annotated near-identity relations[4] was very small: 6% in English and 2% in Catalan. Although clear cases of coreference are undoubtedly the majority, our previous studies on near-identity—on the same AnCora corpus and OntoNotes, which also contains news articles—show that near-identity is not as rare as implied by the figures in Table 2.

Table 3 shows the distribution of `nearidentfeature` tags. The predominant feature by large in the two languages was COMPONENTS, followed by GPE in English, and LOCATION, TIME and OTHER in Catalan. Establishing coreference between plural NPs is indeed a source of major disagreements (Versley, 2008) because it is always problematic to know whether the two sets refer exactly to the same members, as exemplified in (2) above.

These results marked a turning point in our approach to annotating near-identity, also inspired by the results from the online experiment in (Recasens et al., in prep). We concluded that it is not possible to annotate near-identity explicitly because subjects are not aware of it. Our cognitive system tends to perceive reality in terms of simple

[4] Near-identity mentions have an identity degree of 1 or 2.

Identity degree	NP4E		AnCora-CA	
	#	%	#	%
3	5,895	93.8	5,396	98
2	316	5	73	1.3
1	74	1.2	42	0.7

Table 2: Distribution of `identdegree` values (explicit annotation).

Near-identity feature	NP4E		AnCora-CA	
	#	%	#	%
Role	2	0.5	0	0
Representation	0	0	11	4.8
GPE	101	26.1	19	8.3
Components	273	70.4	126	55.0
Location	0	0	25	10.9
Time	6	1.5	23	10.1
Other	6	1.5	25	10.9

Table 3: Distribution of `nearidentfeature` values (explicit annotation).

dichotomies and, to this end, neutralize any minor inconsistency. However, not everyone moves in the same direction. We believe that a factor that contributed to the small percentage of annotated near-identity relations was the fact that annotators were shown the previous coreference annotations. This gave us the key to the right approach: near-identity relations can be inferred from the observed disagreements between annotators.

5. Implicit Near-Identity Annotation

In the light of the previous annotation effort, we moved from explicitly annotating near-identity to implicitly inferring the relations that needed to be annotated as near-identity. The basic idea behind this approach is that different annotators will disagree in labeling a near-identity relation if the only two options that they are given are coreference or non-coreference. So after the same text is annotated by different people in parallel, it is possible to relabel as near-identity those relations that have been annotated as coreferent by some but not all of the annotators. It is an expensive approach, but worth it for a reliable annotation that can be used as the input for machine learning systems to learn this distinction.

Our team of annotators consisted of five linguists. All of them were experienced annotators (not on coreference annotation though), and none of them had participated in the explicit annotation task (Section 4). They were native speakers of Catalan and fluent speakers of English. We asked each of them separately to annotate the same 50K datasets from the NP4E and AnCora-CA corpora that we used for the explicit annotation, but we previously removed all the `entity` tags. The annotation was done in parallel, but they did not share their annotations at any point.

5.1. Annotation guidelines

Unlike the explicit annotation in Section 4, now the annotators were not shown the existing `entity` labels, but it was precisely their job to assign them. They were asked to carry out the classical coreference annotation: for each markable, they had to decide whether it was coreferent or not. All the markables had to be assigned an entity number, and coreferent markables had to be assigned the same entity number as the previous mention.

We did not ask them explicitly to mark near-identity relations, but we still gave them the option of including a `ucoref="yes"` attribute if they were unsure about a coreference relation.

# Annotators	Coreference relations			
	NP4E		AnCora-CA	
	#	%	#	%
5	12,102	46.57	9,287	40.73
4	4,411	16.97	3,314	14.53
3	2,595	9.99	1,989	8.72
2	2,665	10.26	2,081	9.13
1	4,213	16.21	6,131	26.89

Table 4: Raw agreement for pairwise coreference relations.

5.2. Results

Table 4 shows the raw agreement between the five annotators on the annotation of coreference as either yes (identity-of-reference) or no. It reports the number of pairwise coreference relations that were annotated by all five annotators versus those that were annotated by only four, three, two or one annotator. Interestingly, there was full agreement for only half of the relations. If we include those relations annotated by four people and exclude those relations annotated by a single person (which are likely to be a careless error or misunderstanding), then we are left with about 75% of pairwise coreference relations that seem to be clear, and around 25% that are near-identity, for both English and Catalan. It is remarkable that the number of pairwise relations annotated by half (i.e., 2 or 3) of the five annotators is as high as a quarter.

A qualitative analysis of the five parallel annotations (Figure 1) confirms our intuitions that (i) relations annotated by a single annotator represent annotation errors (false positives), (ii) relations annotated by four annotators are to be kept (one annotator missed the link and produced a false negative), and (iii) relations with two or three coreference annotations represent borderline cases of coreference. Figure 1 shows a nice continuum from clear coreference (*the rebels* and *the guerrillas*) to less and less identity (*the standoff* and *the hostage*; *Japan* and *we*). In line with the results from the explicit annotation, the COMPONENTS feature (Section 4.1) seems to be the main one at play, obscuring the boundaries of the referents.

5.3. Merging the annotations

Our approach is new in that we are not interested in quantifying the annotation reliability, but rather in merging the five annotations into a single one and capturing both agreements and disagreements at the same time. Comparing coreference annotations has proved not to be straightforward for either computing inter-annotator agreement (Passonneau, 2004) or evaluating the output of a coreference resolution system (Luo, 2005). One of the biggest difficulties is the dual nature of the coreference problem: we can

5	At an emergency summit in Toronto, the leaders of both nations agreed to push for direct talks with **the rebels**, even though they ruled out **the guerrillas**' non-negotiable demand – freedom for their jailed comrades.
4	**The MRTA** has kept hose considered the best bargaining chips [...] Fujimori told the news conference that Peruvian police would not continue their campaign of provoking **the rebels**.
3	Hashimoto said this would allow Tokyo to play a role in resolving **the standoff** [...] **The hostage crisis** erupted on Dec. 17
2	**Japan** and Peru on Saturday took a tough stand on rebel demands in the Lima hostage crisis [...] "I'm sure that there will be some role **we** can play", Hashimoto said.
1	The leaders of both nations agreed to push for direct talks with the rebels, even though they ruled out **the guerrillas' non-negotiable demand – freedom for their jailed comrades**. [...] Prime Minister Ryutaro Hashimoto supported President Fujimori in his rejection of the MRTA's demand for **the release of the MRTA terrorists currently in incarceration**.

Figure 1: Sample of coreference relations annotated by 5, 4, 3, 2 and 1 annotators.

look at it from a link or from an entity perspective. Therefore, to merge the five annotations we face similar problems with the additional constraint that we need to decide whether a mention is "clear" or "near-identical."

To meet our needs, we designed the algorithm in Figure 2, which consists of two parts: (i) finding the "reliable" links, and (ii) annotating the near-identical mentions. We start with a set of annotations A (its cardinality is represented as $|A|$, which equals 5 in our study) and a set of links L_a that have been annotated by our humans, and we want to obtain the set of final links L_f, i.e., the links that we will use to build the entities of the merged corpus. We first assign a score s_{l_a} to every annotated link l_a as the number of times it has been annotated divided by the total number of annotators. For the final annotation, we only keep the links with a score higher than 0.4 (in the case of our study, links annotated by at least 2 annotators). Using the links in L_f and transitive closure, we obtain the set of entities E, each e consisting of an $|e|$ number of mentions m. For all but the first mention in each entity, its score s_{m_i} equals the sum of all the scores of the links between this mention and its previous mentions divided by the number of links in the entity. The value of the score decides whether the mention is or not near-identical: clear coreferent mentions are those whose score is above 0.5. Mentions with a score between 0.25 and 0.5 are assigned an identity degree equal to 2, and mentions with a score up to 0.25 receive an identity degree equal to 1.

Table 5 compares with Table 2 and shows the distribution of the `identdegree` tags that we obtained after merging the five annotations according to the algorithm in Figure 2. Notice the higher number of near-identity mentions as compared to Table 2: in English, the percentage went from 6% to 12%, and in Catalan it went from 2% to 16%. What is common in both tables is that the number of mentions with identity degree of 2 is larger than that of mentions with identity degree of 1. This suggests that near-identity is usually closer to identity than non-identity, and it can account for the failure of annotators to explicitly mark near-identity.

for all $l_a \epsilon L_a$ **do**
  $s_{l_a} = \frac{|l_a|}{|A|}$
  if $s_{l_a} \geq 0.4$ **then**
    $L_f \leftarrow L_f + l_a$
  end if
end for
for all $e \epsilon E$ **do**
  for $i = |e| \ldots 2$ **do**
    $s_{m_i \epsilon e} = \frac{\sum_{k=1}^{|e|-1} s_{l_a(m_i, m_{i-k})}}{|e|-1}$
    if $s_{m_i} \leq 0.25$ **then**
      $m_i \leftarrow (m_i,$ `identdegree` = 1)
    else if $s_{m_i} \leq 0.5$ **then**
      $m_i \leftarrow (m_i,$ `identdegree` = 2)
    else
      $m_i \leftarrow (m_i,$ `identdegree` = 3)
    end if
  end for
end for

Figure 2: The algorithm for merging coreference annotations and assigning `near-identity` labels.

Identity degree	NP4E		AnCora-CA	
	#	%	#	%
3	6,686	88.19	5,160	83.75
2	823	10.86	909	14.75
1	72	0.95	92	1.50

Table 5: Distribution of `identdegree` values (implicit annotation, after merging the five annotations).

The number of near-identity mentions increased considerably for both languages. As a whole, the results of the implicit annotation seem more reliable than those of the explicit one in the sense that they are closer to our previous analyses, both quantitatively and qualitatively.

We attributed the difference between English and Catalan to the nature of the texts: both corpora consisted of news articles, but we did not want to use translations in order to preserve the native language in each case. We believe that near-identity is a general cognitive relationship, and that the different percentages observed for English and Catalan might be associated with the preference of Romance languages to use a larger number of *unfaithful* anaphors, namely, coreferent NPs whose head is different from that of the previous NP in the chain (Lundquist, 2007).

The two resulting corpora with near-identity tags, NIdent-EN and NIdent-CA, are freely available from `http://clic.ub.edu/corpus/nident`.

6. Conclusion

We presented two annotation studies of near-identity in order to develop the first corpora annotated with near-identity relations. Existing coreferentially annotated corpora tend to lump together, under the label of "coreferent," NPs whose referents are not clearly identical, thus often resulting in inconsistencies and confusing relations. We believe that near-identity cases reduce the performance of automatic coreference resolution systems and therefore their automatic identification could lead to more accurate systems that focus on the clear cases of coreference.

We showed that annotators are not always able to identify linguistic phenomena explicitly, especially subjective ones, and that alternative strategies need to be used. Our approach consisted in using annotation disagreements to arrive at near-identity relations. In this way, we went from 2–6% up to 12–16% of the relations annotated as near-identical, which is a small yet significant number. Considering 12–16% of the relations to be regular coreference cases when they are not is likely to have a negative impact.

There are several avenues for future work, among which testing whether near-identity can be automatically identified, and whether it helps in learning better models of coreference resolution. It would also be interesting to annotate texts other than news articles to compare the types of near-identity relations across domains and genres.

The resulting corpora constitute a useful resource for both developing better coreference resolution systems and for conducting empirical linguistic research to further our understanding of coreference, which involves greater complexity than that assumed by most annotation schemes and coreference resolution systems at present.

7. Acknowledgments

We would like to thank Eduard Hovy for stimulating discussions, and the annotators for their time spent developing the corpora: Celia Alba, Oriol Borrega, Laura Hasler, Blanca Hernández, Difda Monterde, Montse Nofre, and Rita Zaragoza. We also appreciate the reviewers for their constructive feedback.

This work was supported by a Batista i Roca Grant (2010 PBR 00039) and a Beatriu de Pinós scholarship (2010 BP–A 00149) from Generalitat de Catalunya, and the TEXT-Knowledge 2.0 Project (TIN2009-13391-C04-04) from the Spanish Ministry of Science and Innovation.

8. References

Chris Brockett. 2007. Aligning the RTE 2006 Corpus. Technical Report MSR-TR-2007-77, Microsoft Research.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of LREC 2004*, pages 837–840.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of HLT-NAACL 2010*, pages 385–393.

Laura Hasler and Constantin Orasan. 2009. Do coreferential arguments make event mentions coreferential? In *Proceedings of DAARC 2009*, pages 151–163.

Laura Hasler, Constantin Orasan, and Karin Naumann. 2006. NPs for Events: Experiments in coreference annotation. In *Proceedings of LREC 2006*, pages 1167–1172.

Erhard Hinrichs, Sandra Kübler, and Karin Naumann. 2005. A unified representation for morphological, syntactic, semantic and referential annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 Coreference Task Definition – Version 3.0. In *Proceedings of MUC-7*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Proceedings of CoNLL: Shared Task*, pages 28–34.

Lita Lundquist. 2007. Lexical anaphors in Danish and French. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference*, pages 37–48. John Benjamins, Amsterdam, Netherlands.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP 2005*, pages 25–32.

Rebecca Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC 2004*, pages 1503–1506.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC 2008*, pages 1170–1174.

Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*, pages 154–162.

Sameer Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing (ICSC)*, pages 517–526.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of CoNLL: Shared Task*, pages 1–27.

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

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

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, Non-Identity, and Near-Identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Marta Recasens, Liliana Tolchinsky, and M. Antònia Martí.

in prep. Coreference is not always either/or: Psycholinguistic evidence of near-identity.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of LREC 2002*, pages 827–833.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of ANLP 1997*, pages 64–71.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC 2008*, pages 96–101.

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

Kees van Deemter. 2010. *Not Exactly: In Praise of Vagueness*. Oxford University Press, USA.

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

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