A Typology of Near-Identity Relations for Coreference (NIDENT)

Marta Recasens*, Eduard Hovy†, M. Antònia Martí*

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

†USC Information Sciences Institute
4676 Admiralty Way, Marina del Rey CA, United States
hovy@isi.edu

Abstract

The task of coreference resolution requires people or systems to decide when two referring expressions refer to the 'same' entity or event. In real text, this is often a difficult decision because identity is never adequately defined, leading to contradictory treatment of cases in previous work. This paper introduces the concept of 'near-identity', a middle ground category between identity and non-identity, to handle such cases systematically. We present a typology of Near-Identity Relations (NIDENT) that includes fifteen types—grouped under four main families—that capture a wide range of ways in which (near-)coreference relations hold between discourse entities. We validate the theoretical model by annotating a small sample of real data and showing that inter-annotator agreement is high enough for stability ($K = 0.58$, and up to $K = 0.65$ and $K = 0.84$ when leaving out one and two outliers, respectively). This work enables subsequent creation of the first internally consistent language resource of this type through larger annotation efforts.

1. Introduction

The Coreference task requires humans or systems to group together all expressions (typically, NPs) that refer to 'the same thing.' Corpora built in the past by manual annotation include MUC (Hirschman and Chinchor, 1997), ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007), ARRAU (Poesio and Artstein, 2008), AnCora (Recasens and Martí, 2009), PDT (Kučová and Hajičová, 2004), etc. Unfortunately, determining coreference is a highly complex task (Versley, 2008; Poesio and Artstein, 2005), and even these corpora differ on some of the same examples. Compare the two annotations for (1) and (2), where coreferent NPs are printed in italics, and (a) and (b) correspond to the ACE and OntoNotes corpora, respectively.

(1) a. On homecoming night *Postville* feels like Hometown, USA, but a look around this town of 2,000 shows *it*'s become a miniature Ellis Island ... 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.

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

The principal complexity arises when two references denote 'almost' the same thing, either for a single individual (1) or across two groups (2). Such cases are indicative that the binary distinction between coreference (identity) and non-coreference (non-identity) is too limited and fails to account for all the phenomena observed in real data. Instead, coreference relations are better regarded as a continuum with a middle zone of **near-identity** relations. Relations of partial, but not total, identity are the core of the typology presented in this paper.

Near-identity relations resemble 'bridging anaphora' (Clark, 1975) in that they are indirect relations that require the reader to draw an inference, but they differ from typical bridging (Krasavina and Chiarcos, 2007; Nedoluzhko et al., 2009) in that they cannot be said to be relations other than identity. For instance, two NPs that stand in a part-whole semantic relationship can refer to two objects which are related but are not identical, as one is a part of the other —like *the room* and *the ceiling* (3-a)—, in which case we talk of a bridging relation; but it is also possible for them to refer to nearly the same entity (3-b), in which case we must talk of a near-identity relation: although *the United States* in principle denotes the entire US government and *the White House* is restricted to denoting the US president (via metonymy), the discourse in (3-b) employs them as 'almost' coreferent.

(3) a. I looked into *the room*. *The ceiling* was very high.
b. *The United States* has officially restored diplomatic relations with Yugoslavia ... *The White House* said the United States will provide 45 million dollars in food aid to Yugoslavia.

Further evidence for the need of a middle ground relation between identity and non-identity is provided by the fact that Krasavina and Chiarcos (2007) and Nedoluzhko et al. (2009) identify as a source of inter-annotator disagreement cases that can be interpreted as coreference and bridging.

To construct a theoretical model, and before being able to create a corpus, it is necessary to establish how many and what types of near-identity relations exist. The aim of the paper is twofold: (i) to present a typology of coreference relations that includes identity, non-identity, and near-identity

relations; and (ii) to validate the theoretical model by annotating a small sample of real data and showing that inter-annotator agreement is high enough for stability. This work enables subsequent creation of the first internally consistent language resource of this type through larger annotation efforts.

In Section 2. we briefly describe the problem of identity that motivates our project. Section 3. provides the annotation guidelines for Near-Identity Relations (NIDENT). Section 4. outlines a reliability study that establishes inter-annotator agreement at acceptable levels. Finally, we conclude in Section 5.

2. The Problem of Identity

In discourse, we introduce entities or events into a 'reference space' or discourse model by using referring expressions, such as *he*, *John*, *the event*, *then*, etc. These expressions refer to Discourse Entities (DEs), which are mental representations existing in our minds (Karttunen, 1976; Webber, 1979). Though they often correlate with entities and events in the external world, they need not. It is upon DEs, not the external referring expressions, that coreference operates, unlike typical definitions of coreference in the NLP literature that talk about real-world entities (Finkel and Manning, 2008; Ng, 2009).

To identify a DE, the speaker provides one or more of its attributes, the most basic one being the name of the concept, entity, or event. When DEs are mentioned multiple times throughout a discourse, new attributes might be introduced, old attributes might be omitted or changed, etc. The speaker states a series of attributes that the hearer is able to recognize or know as (supposedly) true of the DE (at that time), enough to pick it out uniquely. The problem of coreference is determining when expressions in a discourse refer to the 'same' DE, and when they introduce a new (albeit possibly very closely related) one. The different possible ways of coreferring pertain to different identity criteria.

Since DEs are anchored in discourse, so is coreference. We cannot, for instance, make a general statement that *Barack Obama* and *Obama* are coreferent without analyzing the discourse in which they appear, i.e., the DEs they refer to. Compare (4-a) and (4-b).

(4) a. The Clinton campaign is circulating a fake photo of *Barack Obama in Muslim clothes* to damage his reputation. In the photo, *Obama* wears a long sari-like garment.

b. The Clinton campaign is circulating a fake photo of *Barack Obama in Muslim clothes* to damage his reputation, but *Obama* never wore Muslim clothes.

The two italicized expressions in (4-a) corefer as they both refer to the DE corresponding to Barack Obama as represented in a specific photo. The very same expressions in (4-b), however, stand in a more complex relation: whereas the first NP refers to a photo representation of Barack Obama, the second NP refers to the actual person of Barack Obama. In this regard, then, they are not identical, yet they are in the sense that the former stands for the latter. Notice that we could even replace *Obama* for *he* in (4-b). This dilemma between identity and non-identity is the reason why the concept of 'near-identity' (akin to Geach's (1967) 'relative identity') needs to enter the field of coreference resolution. Currently, automatic coreference resolution systems make no distinction between (4-a) and (4-b), which can lead to erroneous or contradictory conclusions. A user searching for information about Barack Obama would learn from an information extraction system that Obama wears a long sari-like garment and that he never wore Muslim clothes.

Another important dimension that causes a conflict between identity and non-identity is time. The attributes of an entity change through time, thus the problems with (1). Similarly, the different roles an entity can play make it hard to analyze a discourse in terms of only coreference or non-coreference. In (5),[1] the reader is forced to identify *the artist* and *the man* separately; identifying *the one* or *the other* with the central person would result in an incoherent discourse.

(5) You cannot read Cyril Connolly for very long without wanting to acquire —and then developing— a relationship with the personality of the man himself. [...] With Connolly there is a marked difference and the difference is that *the artist* and *the man* are so conjoined and intermingled that you cannot savour *the one* without *the other* and vice versa.

The shortcomings of a coreference definition grounded on real-world reference and on a binary conception of identity became evident when annotating the AnCora corpus with coreference information (Recasens and Martí, 2009) as well as when training coreference resolution systems on the OntoNotes and ACE corpora (Doddington et al., 2004; Pradhan et al., 2007). This led us to develop the NIDENT typology that we present in the next section.

3. A Typology of (Near-)Identity Relations

This section presents NIDENT, a typology of coreference relations that contemplates relations of partial, but not total, identity by including near-identity relations. Table 1 gives a schematic summary. We followed a bottom-up approach consisting in first extracting problematic coreference relations[2] from real data, and then analyzing them and defining a set of coreference types that accounted for the observed differences.

1. **Non-Identity**. The two NPs point to two different DEs. Even if they share any feature, they are not 'the same thing.'

   President Samaranch sent *a letter* to Sydney in which he asked for information. *A similar missive* has also been received by all the candidate cities to host the Olympic Games of 1996.

2. **Identity**. The two NPs point to the same DE (i.e., they have the same set of attributes, as far as one can tell). They are (almost certainly) 'the same thing.'

[1] Example taken from William Boyd's *Bamboo* (2005).

[2] By *problematic* we mean those cases that involved disagreements between the annotators or that could be argued either way —coreferent or non-coreferent— according to the authors.

Relation	Type	Subtype
1. Non-Identity		
2. Identity		
3. Near-Identity	A. Name metonymy	a. Role b. Location c. Organization d. Information realization e. Representation f. Other
	B. Meronymy	a. Part·Whole b. Stuff·Object c. Set·Set
	C. Class	a. More specific b. More general
	D. Spatio-temporal function	a. Place b. Time c. Numerical function d. Role function

Table 1: The NIDENT typology.

> It began when *a Hasidic Jewish family* bought one of the town's two meat-packing plants 13 years ago. First *they* brought in other Hasidic Jews, then Mexicans, Palestinians, Ukrainians.

Sometimes, the second NP provides more information about the DE, or selects a subset of characteristics of the DE and ignores the rest. If this introduces no opposition or contradiction between the new attributes and the ones already known, then the new NP refers to the same DE.

> While the Cubans from Miami and *Cuba* fight for the custody of little Elián González, US companies do not take their eyes off the juicy market that *the Caribbean pearl* has become.

Referring to Cuba as a *pearl* stresses what accounts for the interest of US companies in the island while leaving out Cuba's poverty. But we can still talk of Identity as there is no aim to draw any opposition.

3. **Near-Identity**. If two NPs are neither Identical nor Non-Identical, they are partially the same in that they share most of the important characteristics, but differ in at least one crucial dimension. According to the dimension of contradiction or opposition, the following types and subtypes can be distinguished.

   3.A **Name metonymy.** Since many entities in the world are complex, multifaceted, it is possible to refer to the same entity via different facets. For example, a company produces a product, is headquartered in a location, has a president, etc. Under Name metonymy, a proper noun (a name) appears first, and a subsequent NP refers to some facet of the DE. Since the specific facets available depend on the type of entity under consideration, there are a great many possibilities. Nonetheless, certain facets occur frequently enough that we create specially named subtypes for them. All the remainder we group as the OTHER category (3.A.f). The subtypes are the following.

   3.A.a. ROLE. A specific role or function performed by a human, animal or object, is distinguished from their other facets.

   > "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.

   The *actor* and *father* pertain to two different roles of the same individual *Gassman*. The opposition here pertains to the typical activities of Gassman: actor-like actions or father-like ones. Presumably, they are not the same, given the opposition expressed in the citation.

   3.A.b. LOCATION. As a metaconcept, the name of a location can be used indiscriminately to describe facet(s) such as the physical place, the place associated with a (political) organization, the population living in that location, the ruling government, an affiliated organization, an event celebrated at that location, etc.

   > The Jordan authorities arrested, on arriving in *Iraq*, an Italian pilot who violated the air embargo to *this country*.

   LOCATION can co-occur with PART·WHOLE (3.B.a). Metonymy wins over meronymy, so 3.A.b is preferred when they co-occur.

   3.A.c. ORGANIZATION. As a metaconcept, the name of a company or other social organization can be used indiscriminately to describe facet(s) such as the legal organization itself, the facility that houses the organization or one of its branches, the company shares, a product manufactured by the company, etc.

   > The strategy has been a popular one for *McDonalds* . . . It's a very wise move on for them because if they would have *only just original McDonalds*, I don't think they would have done so great.

   This type is chosen if it co-occurs with 3.B.a.

3.A.d. INFORMATIONAL REALIZATION. A DE corresponding to an informational object (e.g., story, law, review, etc.) can be split according to the format in which the information is presented or manifested (e.g., book, movie, speech, etc.). The content, however, is shared by all the DEs.

> She hasn't seen *Gone with the Wind*, but she's read *it*.

The two DEs pertain to a film and a book, which are not identical, though their content —the story— is.

3.A.e. REPRESENTATION. One NP is a representation of the other —as in a picture or a starring of a person, or a toy replica of a real object. The representation can also be of a more abstract kind, like one's mental conceptualization of an object. One NP corresponds to the thing represented; the other, to the element that represents it.

> We stand staring at two paintings of *Queen Elizabeth*. In the one on the left, *she* is dressed as Empress of India. In the one on the right, *she* is dressed in an elegant blue gown.

Each painted queen is a different DE even though they are all the same queen: they can coexist, and each one is dressed differently.

3.A.f. OTHER. Since metonymic readings are potentially open-ended and can be invented dynamically, this subtype captures any other case.

> *Chevrolet* is a brand of automobile produced by General Motors Company. *It* is feminine because of its sound.

3.B **Meronymy**. A meronym (a constituent part or a member of a composite entity) is used to refer to the whole or set (Chaffin et al., 1988).

3.B.a. PART·WHOLE. One NP mentions a part to refer to the whole expressed by the other NP. The two NPs can be interpreted as referring to nearly the same DE because one expresses a functionally very relevant part of the whole it belongs to. The whole is composed of different, functionally distinct, parts (e.g., *the engine* and *the car*), organized into some kind of patterned organization or structure.

> Bangladesh Prime Minister Hasina and *President Clinton* expressed the hope that this trend will continue ... Both *the US government* and American businesses welcomed the willingness of Bangladesh.

Here *President Clinton* is seen as a functioning part of the entire *US government*. Since the president acts as the spokesperson of the government, they are often used indistinctively as near-identical. But clearly, there are other, quite different, parts in the government.

3.B.b. STUFF·OBJECT. One NP expresses the constituent material of the other NP. Unlike components, the stuff of which a thing is made cannot be separated from the object.

> The City Council approved legislation prohibiting selling *alcoholic drinks* during night hours ... Bars not officially categorized as bars will not be allowed to sell *alcohol*.

Given that the most relevant component of alcoholic drinks is alcohol, this can be used to refer to the drinks. Bars do not sell alcohol by itself, but alcoholic drinks.

3.B.c. SET·SET. The two NPs denote two largely overlapping sets. Since each set is not clearly bounded, the reader intuitively interprets the two sets as near-identical even though they might not correspond to exactly the same collection of individuals. Unlike 3.B.a, the collection consists of repeated, closely similar, members, and the members are not required to perform a particular function distinct from one another. This is the preferred type whenever a strict identity relation between the two plural NPs is dubious (no total overlapping).

> Last night in Tel Aviv, *Jews* attacked a restaurant that employs Palestinians "*we* want war," *the crowd* chanted.

For the purposes of the discourse, the three NPs *Jews*, *we* and *the crowd* can be considered to refer to the same set, although not every Jew that attacked the restaurant might have chanted "we want war," or one of the individuals who chanted "we want war" might not have been Jewish.

SET·SET shows an interesting pattern when the first NP denotes a well-established (often named) set and the second NP denotes the sum of its members.

> The administration of *the European Union* has demonstrated neither an excellent foresight capacity nor a capacity of resolution ... It has not taken into account the characteristics of *the countries in the EU*, which have large differences in their systems of medical practice.

The second NP refers to the EU via the collection of its members, although theoretically we know that the organization is greater than the sum of the parts (since the EU has its own governing authorities, its own resources, its own agenda, etc.).

3.C **Class**. The two NPs share the type (is-a relationship), but they stand in a different position in the categorical hierarchy so that one can be viewed as more general or specific to the other.

3.C.a. MORE SPECIFIC. The second NP expresses a class more specific than that expressed by the first NP so that the predication affects only a subset of the class.

It shows that *attackers*, like *the two who killed 13 people at Columbine High School last year in Colorado*, come from a variety of family and ethnic backgrounds.

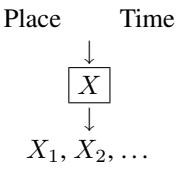
3.C.b. MORE GENERAL. Similar to 3.C.a but in the opposite direction: The second NP goes one level up in the conceptualization hierarchy by expressing a more general class.

Diego looked for information about *his character* in the novel forgetting that Saramago does not usually describe *them*.

The introduction of Diego's character allows the text to generalize and talk about Saramago's characters as a class, but *his character* is only one of *them*.

If it is hard to decide between 3.C.a and 3.C.b, then 3.B.c is probably the solution.

3.D **Spatio-temporal function**. The DE is split based on different values for its spatial or temporal characteristics: it is the 'same' entity or event but realized in another location or time. One can imagine a function or box (the superconcept of the DE) to which a place/time argument is provided (by clicking a place and/or a time), obtaining a different instance of the DE in each case.



Often, the sentence includes a temporal or locative phrase (e.g., *in 1999*, *old*, etc.) that provides the argument. This function assumes 3.C, but 3.D is preferred since it is more specific. We differentiate the following four subtypes.

3.D.a. PLACE. The same DE is instantiated in different physical locations, each time resulting in a different DE due to the change in the spatial feature. It is possible for them to coexist but not in the same place.

New York's New Year's Eve is one of the most widely attended parties in the world ... Celebrating *it* in the Southern Hemisphere is always memorable, especially for those of us in the Northern Hemisphere.

Although the two NPs refer to the same celebration, New Year, the celebration in New York is clearly not the same as the celebration in the Southern Hemisphere.

3.D.b. TIME. Similar to 3.D.a but the split into different DEs is due to a change of the temporal value. It sees a physical object as a function from time to a portion of space, a slice of the object's history. Thus, it is not possible for the temporally-different DEs to coexist.

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.

3.D.c. NUMERICAL FUNCTION. The two NPs refer to the same function (e.g., price, age, rate, etc.) but have different numerical value due to a change in time or a change in space. Although 3.D.a or 3.D.b might apply, 3.D.c is more specific.

At 8, *the temperature* rose to 99º. This morning *it* was 85º.

The two DEs differ in that the value of the first is 99 degrees while the value of the second is 85 degrees, but they are near-identical in that they refer to the same temperature function.

3.D.d. ROLE FUNCTION. The two NPs refer to the same role (e.g., president, director, etc.) but is filled by a different person due to a change in time or space. Although 3.D.a or 3.D.b might apply, 3.D.d is more specific.

In France, *the president* is elected for a term of seven years, while in the United States *he* is elected for a term of four years.

The two DEs differ in that the value of the first is the president of France (e.g., Sarkozy) while the value of the second is the president of the US (e.g., Obama), but they are near-identical in that they refer to the same role.

For a more complete description of the theory underlying this typology, we address the reader to Recasens et al. (in prep).

4. Stability Study

In the process of establishing how many and what types of near-identity relations exist, we carried out three annotation rounds on a sample of naturally occurring data to help identify weaknesses of previous typology versions and refine the theoretical model.

4.1. Method

Participants Six paid subjects participated in the study: four undergraduate students and two authors of this paper. Although the undergraduates were not linguistics students, they were familiar with annotation tasks requiring semantic awareness, but had not worked on coreference before.

Materials A total of 60 text excerpts were selected from three electronic corpora —ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007) and AnCora (Recasens and Martí, 2009)— as well as from the Web, a television show, and real conversation. The excerpts were divided in three rounds of 20, each including examples of the different identity types in different proportions so that annotators could not reason by elimination or the like. To the same effect, each round varied the proportions (see Table 2), yet the largest number of examples always was Near-Identity, as we were mainly interested in evaluating to what extent coders discriminated among the fine-grained subtypes.

Round	Non-Identity	Identity	Near-Identity
First	4%	40%	56%
Second	8%	28%	64%
Third	7%	12%	81%

Table 2: Proportion of types in the stability study.

[*Halle Berry speaking*] At an early age my daughter would recognize [me]$_1$ in the photo ... she's over my shoulder and I hear her say ["Mama, mama"]$_2$, and I knew "Oh, she saw that cover, that's cute." And this woman behind her was sort of cooing with her, and I heard the woman say "Oh, no, honey, [that]$_3$'s not your mama, that's Halle Berry."

NP$_a$	NP$_b$	Identity Types	Note
1	2		
1	3		
2	3		

Figure 1: Sample item from the stability study.

In each excerpt, two or more NPs were marked with square brackets and were given a subscript ID number. Apart from the set of 20 excerpts, annotators were given an answer sheet where all the possible combinatorial pairs between the marked NPs were listed. See a sample item from the annotation task in Figure 1. The first 20 excerpts included 78 pairs to be analyzed; the second, 53, and the third, 43.[3]

Procedure The task required coders to read the annotation guidelines and classify the selected pairs of NPs in each excerpt according to the (near-)identity relation(s) that obtained between them by filling in the answer sheet. They had to assign one or more, but at least one, class to each pair of NPs, indicating the corresponding type and subtype identifiers. They were asked to specify all the possible (sub)types for underspecified pronouns and genuinely ambiguous NPs that accepted multiple interpretations, and to make a note of comments, doubts or remarks they had.
The three groups of 20 excerpts were annotated in three separate rounds, spread out over a span of four weeks. In each round, a different version of the annotation guidelines was used, since the typology underwent substantial revision —in a decreasing manner— after completing each round. The typology presented in Section 3. corresponds to the final version that was used in the last round.

4.2. Results and Discussion

Inter-coder agreement was measured with Fleiss's kappa (Fleiss, 1981), computed with the R package *irr*, version 0.82 (Gamer et al., 2009). Statistical significance was tested with a kappa z-test provided by the same package. Kappa calculates the degree of agreement in classification over that which would be expected by chance and its values range between -1 and 1, where 1 signifies perfect agreement, 0 signifies no difference from chance agreement, and negative values signify that agreement is weaker than expected by chance. Typically, a kappa value of at least 0.60 is required. For the cases in which a coder gave multiple classes as answer, the one showing the highest agreement was used for computing kappa.

The first round of 20 texts, which served as a practice round, only obtained $K = 0.32$. More importantly, it revealed serious shortcomings of the first version of the typology. In this regard, the comments and notes included by the coders in the answer sheet were very helpful. Cases for which four —or even five— different types were suggested, especially those that included both IDENTITY and NON-IDENTITY answers, manifested the difficulty and subjectivity of the task. Some cases could be addressed by including an additional type or restricting the scope of an existing one. Second, multiple answers given by the same annotator (the highest number of types being three) were in general indicative that the definitions provided in the guidelines were not specific enough and more criteria had to be included to choose between the different types. In some cases, however, there were problems of overlapping, i.e., two types were equally valid for the same relation, and we decided to merge them. Finally, broad types without identifying force were removed.

As a result of the major revision of the guidelines after the first round, the agreement of the second set of 20 texts reached $K = 0.54$. In contrast with the previous round, the answers were not so spread over different types: a larger number of pairs obtained total agreement, and low disagreements were concentrated in a few types. To address this, a solution was found in setting clear preferences in the guidelines for cases when it was possible for two types to co-occur with no contradiction, like LOCATION and PART·WHOLE. A large number of isolated (5-to-1) answers made us consider the possible presence of outliers, and we detected two. If agreement was computed between the other five coders, we obtained a considerable improvement resulting in $K = 0.63$; between the other four coders, $K = 0.71$.

The third and final round obtained a further improvement in agreement: $K = 0.58$, and up to $K = 0.65$ and $K = 0.84$ leaving out the one and two outliers, respectively. Table 3 details the per-class and overall K of this last round. The changes introduced in the typology after the second round were small compared with the revision we had undertaken after the first round. Basically, we improved the guidelines by adding some clarifications and commenting all the examples. Setting the preferences of some types over others also contributed in the overall improvement. Nevertheless, the (near-)identity task is difficult and requires a mature sensitivity to language that not all coders had, as revealed by the presence of outliers.

To exemplify a case of 50%-50% disagreement, consider (6), which obtained four SET·SET answers, four TIME, and one NON-IDENTITY. Proponents of the first saw in *the people* and *they* a relation between two underspecified plurals, proponents of the second saw a clear temporal function at play, distinguishing the people from the past and the people from today; the last proponent failed to see any of the two previous near-identity relations and interpreted them as two groups with no identity at all.

[3] We address the reader to Recasens et al. (in prep) for the entire collection of texts as they make evident the difficulty of the task. We are happy to send it to interested requesters.

Relation	Type	Subtype	*K*	z	p-value
1. Non-Identity			0.89	22.64	0.00
2. Identity			0.30	7.55	0.00
3. Near-Identity	A. Name metonymy	a. Role	-0.00	-0.10	0.92
		b. Location	0.87	22.01	0.00
		c. Organization	0.48	12.09	0.00
		d. Information realization	0.49	12.54	0.00
		e. Representation	0.59	15.08	0.00
		f. Other	0.59	15.08	0.00
	B. Meronymy	a. Part·Whole	-0.00	-0.10	0.92
		b. Stuff·Object	0.80	20.22	0.00
		c. Set·Set	0.73	18.44	0.00
	C. Class	a. More specific	0.39	9.80	0.00
		b. More general	0.38	9.61	0.00
	D. Spatio-temporal function	a. Place	0.67	16.90	0.00
		b. Time	0.70	17.70	0.00
		c. Numerical function			
		d. Role function	-0.01	-0.20	0.84
Total			0.58	39.50	0.00

Table 3: Final results of the stability study (Round 3).

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

The general picture that emerges is that of a continuum: there are prototype examples that clearly illustrate each relation plus a whole range of intermediate cases with fuzzy boundaries. This is the reason why we did not merge any other types. The fact, however, that each coder showed their own preference for one of the two possible types could indicate that a 'coreference bias' —a preference for a specific perspective when establishing coreference links— predominates in every person. The psycholinguistic study of this and other aspects remains as future work.

5. Conclusion

We have motivated the need for a middle ground category between identity and non-identity in the coreference task by showing examples from naturally occurring data that cannot be accounted for under a binary coreference definition. The NIDENT typology includes fifteen near-identity types —grouped under four main families— that capture a wide range of ways in which (near-)coreference relations hold between discourse entities. We see it as an open, rather than definitive, typology as there are many (creative) ways in which discourse entities can be connected in a near-identical fashion, but we believe that it provides main axes along which near-identity can manifest itself in discourse.

Support for NIDENT came from an inter-annotator agreement study, which showed that agreement increased over three consecutive rounds, achieving $K = 0.58$, and up to $K = 0.65$ and $K = 0.84$ when leaving out one and two outliers, respectively. The assumption of a near-identity category in the processing of coreference relations has consequences not only for computational linguistics but from theoretical to cognitive linguistics. We have hinted at the potential future work in the field of psycholinguistics.

6. Acknowledgments

We are grateful to Jerry Hobbs for his insight and to the annotators: David Halpern, Peggy Ho, Justin James, and Rita Zaragoza.

This work was supported in part by FPU Grant AP2006-00994 from the Spanish Ministry of Education, and TEXT-MESS 2.0 (TIN2009-13391-C04-04) Project.

7. References

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

Herbert H. Clark. 1975. Bridging. In R. C. Schank and B. L. Nash-Webber, editors, *Theoretical Issues in Natural Language Processing*, pages 169–174. Association for Computing Machinery, New York.

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*, Lisbon.

Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-HLT 2008*, pages 45–48, Columbus, Ohio.

J.L. Fleiss. 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 2nd edition.

Matthias Gamer, Jim Lemon, and Ian Fellows, 2009. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.82.

Peter Geach. 1967. Identity. *Review of Metaphysics*, 21:3–12.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 Coreference Task Definition – Version 3.0, July.

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

Olga Krasavina and Christian Chiarcos. 2007. PoCoS – Postdam Coreference Scheme. In *Proceedings of the First Linguistic Annotation Workshop (LAW I)*, pages 156–163, Prague.

Lucie Kučová and Eva Hajičová. 2004. Coreferential relations in the Prague Dependency Treebank. In *Proceedings of DAARC 2004*, pages 97–102, San Miguel, Azores.

Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 108–111, Singapore.

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

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II*, Ann Arbor, Michigan.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC 2008*, Marrakech.

Sameer S. 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-07)*, pages 517–526, Washington, DC.

Marta Recasens and M. Antònia Martí. 2009. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, DOI 10.1007/s10579-009-9108-x.

Marta Recasens, Eduard Hovy, and M. Antònia Martí, in prep. *Identity, non-identity, and near-identity: Addressing the complexity of coreference*.

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

Bonnie Lynn Webber. 1979. *A Formal Approach to Discourse Anaphora*. Garland Press, New York.