# Coreference is not always either/or: Psycholinguistic evidence of near-identity

Marta Recasens, Stanford University

Coreference has been traditionally defined dichotomously as identity-of-reference or non-identity-of-reference. Accordingly, the linguistic annotations of coreferentially annotated corpora (Doddington et al., 2004; Hirschman & Chinchor, 1997; Pradhan et al., 2007) identify the NPs that refer to the *same* entity from those that refer to *different* entities. However, we have pointed out in previous work (Recasens, Hovy, & Martí, 2011) that there are examples in real data that fall under neither coreference nor non-coreference, as two referents can be just *nearly* identical. This is evidenced by the fact that *the old Postville* is annotated contradictorily by the ACE (1-a) and OntoNotes (1-b) corpora (mentions annotated as coreferent are marked in italics).

(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.

Adding to previous work on referential ambiguity and vagueness (van Deemter, 2010; Poesio, Sturt, Artstein, & Filik, 2006; Versley, 2008), we provide psycholinguistic evidence of the near-identity category, gaining further insight into the nature of near-identity relations. More specifically, we present a three-task experiment on the interpretation of the identity relationship between the referents of noun phrases in English and in Catalan. The experiment collected judgments from 104 subjects (70 English speakers and 34 Catalan speakers) who were asked to rate the referential sameness of two NPs in (i) a two-way choice task, (ii) a four-point scale, and (iii) a multiple-choice question type.

The results show that whereas some referential relations are classified as either Identity or Non-identity by the majority of subjects, there is a third class of relations for which the judgments are split between Identity and Non-identity. This evidence supports the conclusion that it does not suffice to distinguish between identity-of-reference and non-identity-of-reference, but that it is psychologically plausible to assume a middle-ground category of near-identity to include those referential relations that part of the subjects classify as Identity and the other part as Non-identity. In addition, it emerges that such relations have a longer response time, which reveals that near-identity has a higher processing difficulty. The fact that this is true for both English and Catalan points toward the crosslinguistic nature of near-identical referents.

# References

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 LREC 2004* (pp. 837–840). Lisbon.

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

Poesio, M., Sturt, P., Artstein, R., & Filik, R. (2006). Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes: a multidisciplinary journal*, *42*, 157–175.

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

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

van Deemter, K. (2010). *Not exactly: In praise of vagueness.* Oxford University Press, USA.

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