

On Paraphrase and Coreference

Marta Recasens*

University of Barcelona

Marta Vila**

University of Barcelona

By providing a better understanding of paraphrase and coreference in terms of similarities and differences in their linguistic nature, this article delimits what the focus of paraphrase extraction and coreference resolution tasks should be, and to what extent they can help each other. We argue for the relevance of this discussion to Natural Language Processing.

1. Introduction

Paraphrase extraction¹ and coreference resolution have applications in Question Answering, Information Extraction, Machine Translation, and so forth. Paraphrase pairs might be coreferential, and coreference relations are sometimes paraphrases. The two overlap considerably (Hirst 1981), but their definitions make them significantly different in essence: Paraphrasing concerns meaning, whereas coreference is about discourse referents. Thus, they do not always coincide. In the following example, *b* and *d* are both coreferent and paraphrastic, whereas *a*, *c*, *e*, *f*, and *h* are coreferent but not paraphrastic, and *g* and *i* are paraphrastic but not coreferent.

- (1) [Tony]_a went to see [the ophthalmologist]_b and got [his]_c eyes checked. [The eye doctor]_d told [him]_e that [his]_f [cataracts]_g were getting worse. [His]_h mother also suffered from [cloudy vision]_i.

The discourse model built for Example (1) contains six entities (i.e., Tony, the eye doctor, Tony's eyes, Tony's cataracts, Tony's mother, cataracts). Because *a*, *c*, *e*, *f*, and *h* all point to Tony, we say that they are coreferent. In contrast, in paraphrasing, we do not need to build a discourse entity to state that *g* and *i* are paraphrase pairs; we restrict ourselves to semantic content and this is why we check for sameness of meaning between *cataracts* and *cloudy vision* alone, regardless of whether they are a referential unit in a discourse. Despite the differences, it is possible for paraphrasing and coreference to co-occur, as in the case of *b* and *d*.

NLP components dealing with paraphrasing and coreference seem to have great potential to improve understanding and generation systems. As a result, they have been the focus of a large amount of work in the past couple of decades (see the surveys by

* CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: mrecasens@ub.edu.

** CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: marta.vila@ub.edu.

1 Recognition, extraction, and generation are all paraphrase-related tasks. We will center ourselves on paraphrase extraction, as this is the task in which paraphrase and coreference resolution mainly overlap.

Androutsopoulos and Malakasiotis [2010], Madnani and Dorr [2010], Ng [2010], and Poesio and Versley [2009]). Before computational linguistics, coreference had not been studied on its own from a purely linguistic perspective but was indirectly mentioned in the study of pronouns. Although there have been some linguistic works that consider paraphrasing, they do not fully respond to the needs of paraphrasing from a computational perspective.

This article discusses the similarities between paraphrase and coreference in order to point out the distinguishing factors that make paraphrase extraction and coreference resolution two separate yet related tasks. This is illustrated with examples extracted/adapted from different sources (Dras 1999; Doddington et al. 2004; Dolan, Brockett, and Quirk 2005; Recasens and Martí 2010; Vila et al. 2010) and our own. Apart from providing a better understanding of these tasks, we point out ways in which they can mutually benefit, which can shed light on future research.

2. Converging and Diverging Points

This section explores the overlapping relationship between paraphrase and coreference, highlighting the most relevant aspects that they have in common as well as those that distinguish them. They are both sameness relations (Section 2.2), but one is between meanings and the other between referents (Section 2.1). In terms of linguistic units, coreference is mainly restricted to noun phrases (NPs), whereas paraphrasing goes beyond and includes word-, phrase- and sentence-level expressions (Section 2.3). One final diverging point is the role they (might) play in discourse (Section 2.4).

2.1 Meaning and Reference

The two dimensions that are the focus of paraphrasing and coreference are meaning and reference, respectively. Traditionally, paraphrase is defined as the relation between two expressions that have the same *meaning* (i.e., they evoke the same mental concept), whereas coreference is defined as the relation between two expressions that have the same *referent* in the discourse (i.e., they point to the same entity). We follow Karttunen (1976) and talk of “discourse referents” instead of “real-world referents.”

In Table 1, the italicized pairs in cells (1,1) and (2,1) are both paraphrastic but they only corefer in (1,1). We cannot decide on (non-)coreference in (2,1) as we need a discourse to first assign a referent. In contrast, we can make paraphrasing judgments

Table 1
Paraphrase–coreference matrix.

	✓	Paraphrase	✗
Coreference	✓	(1,1) Tony went to see <i>the ophthalmologist</i> and got his eyes checked. <i>The eye doctor</i> told him ...	(1,2) Tony went to see the ophthalmologist and got <i>his</i> eyes checked.
	✗	(2,1) <i>ophthalmologist eye doctor</i>	(2,2) <i>His cataracts</i> were getting worse. <i>His mother</i> also suffered from cloudy vision.

without taking discourse into consideration. Pairs like the one in cell (1,2) are only coreferent but not paraphrases because the proper noun *Tony* and the pronoun *his* have reference but no meaning. Lastly, neither phenomenon is observed in cell (2,2).

2.2 Sameness

Paraphrasing and coreference are usually defined as sameness relations: Two expressions that have the *same meaning* are paraphrastic, and two expressions that refer to the *same entity* in a discourse are coreferent. The concept of *sameness* is usually taken for granted and left unexplained, but establishing sameness is not straightforward. A strict interpretation of the concept makes sameness relations only possible in logic and mathematics, whereas a sloppy interpretation makes the definition too vague. In paraphrasing, if the loss of *at the city* in Example (2b) is not considered to be relevant, Examples (2a) and (2b) are paraphrases; but if it is considered to be relevant, then they are not. It depends on where we draw the boundaries of what is accepted as the “same” meaning.

- (2) a. The waterlogged conditions that ruled out play yesterday still prevailed *at the city* this morning.
- b. The waterlogged conditions that ruled out play yesterday still prevailed this morning.
- (3) On homecoming night *Postville* feels like Hometown, USA ...For those who prefer *the old Postville*, Mayor John Hyman has a simple answer.

Similarly, with respect to coreference (3), whether *Postville* and *the old Postville* in Example 3 are or are not the same entity depends on the granularity of the discourse. On a sloppy reading, one can assume that because *Postville* refers to the same spatial coordinates, it is the same town. On a strict reading, in contrast, drawing a distinction between the town as it was at two different moments in time results in two different entities: the old *Postville* versus the present-day *Postville*. They are not the same in that features have changed from the former to the latter.

The concept of sameness in paraphrasing has been questioned on many occasions. If we understood “same meaning” in the strictest sense, a large number of paraphrases would be ruled out. Thus, some authors argue for a looser definition of paraphrasing. Bhagat (2009), for instance, talks about “quasi-paraphrases” as “sentences or phrases that convey approximately the same meaning.” Milićević (2007) draws a distinction between “exact” and “approximate” paraphrases. Finally, Fuchs (1994) prefers to use the notion of “equivalence” to “identity” on the grounds that the former allows for the existence of some semantic differences between the paraphrase pairs. The concept of identity in coreference, however, has hardly been questioned, as prototypical examples appear to be straightforward (e.g., *Barack Obama* and *Obama* and *he*). Only recently have Recasens, Hovy, and Martí (2010) pointed out the need for talking about “near-identity” relations in order to account for cases such as Example (3), proposing a typology of such relations.

2.3 Linguistic Units

Another axis of comparison between paraphrase and coreference concerns the types of linguistic units involved in each relation. Paraphrase can hold between different

linguistic units, from morphemes to full texts, although the most attention has been paid to word-level paraphrase (*kid* and *child* in Example (4)), phrase-level paraphrase (*cried* and *burst into tears* in Example (4)), and sentence-level paraphrase (the two sentences in Example (4)).

- (4) a. The kid cried.
b. The child burst into tears.

In contrast, coreference is more restricted in that the majority of relations occur at the phrasal level, especially between NPs. This explains why this has been the largest focus so far, although prepositional and adverbial phrases are also possible yet less frequent, as well as clauses or sentences. Coreference relations occur indistinctively between pronouns, proper nouns, and full NPs that are *referential*, namely, that have discourse referents. For this reason, pleonastic pronouns, nominal predicates, and appositives cannot enter into coreference relations. The first do not refer to any entity but are syntactically required; the last two express properties of an entity rather than introduce a new one. But this is an issue ignored by the corpora annotated for the MUC and ACE programs (Hirschman and Chinchro 1997; Doddington et al. 2004), hence the criticism by van Deemter and Kibble (2000).

In the case of paraphrasing, it is linguistic expressions that lack meaning (i.e., pronouns and proper nouns) that should not be treated as members of a paraphrase pair on their own (Example (5a)) because paraphrase is only possible between meaningful units. This issue, however, takes on another dimension when seen at the sentence level. The sentences in Example (5b) can be said to be paraphrases because they themselves contain the antecedent of the pronouns *I* and *he*.

- (5) a. (i) A. Jiménez
(ii) I
b. (i) The Atlético de Madrid goalkeeper, A. Jiménez, yesterday realized one of his dreams by defeating Barcelona: "I had never beaten Barcelona."
(ii) The Atlético de Madrid goalkeeper, A. Jiménez, yesterday realized one of his dreams by defeating Barcelona, and said that he had never beaten Barcelona.

In Example (5b), *A. Jiménez* and *I/he* continue not being paraphrastic. Polysemic, underspecified, and metaphoric words show a slightly different behavior. It is not possible to establish paraphrase between them when they are deprived of context (Callison-Burch 2007, Chapter 4). In Example (6a), *police officers* could be patrol police officers, and *investigators* could be university researchers. However, once they are embedded in a disambiguating context that fills them semantically, as in Example (6b), then paraphrase can be established between *police officers* and *investigators*.

- (6) a. (i) Police officers
(ii) Investigators
b. (i) *Police officers* searched 11 stores in Barcelona.
(ii) The *investigators* conducted numerous interviews with the victim.

As a final remark, and in accordance with the approach by Fuchs (1994), we consider Example (7)–like paraphrases that Fujita (2005) and Milićević (2007) call, respectively,

“referential” and “cognitive” to be best treated as coreference rather than paraphrase, because they only rely on referential identity in a discourse.

- (7) a. They got married *last year*.
 b. They got married *in 2004*.

2.4 Discourse Function

A further difference between paraphrasing and coreference concerns their degree of dependency on discourse. Given that coreference establishes sameness relations between the entities that populate a discourse (i.e., discourse referents), it is a linguistic phenomenon whose dependency on discourse is much stronger than paraphrasing. Thus, the latter can be approached from a discursive or a non-discursive perspective, which in turn allows for a distinction between reformulative paraphrasing (Example (8)) and non-reformulative paraphrasing (Example (9)).

- (8) Speaker 1: Then they also diagnosed *a hemolytic–uremic syndrome*.
 Speaker 2: What’s that?
 Speaker 1: *Renal insufficiency, in the kidneys*.
- (9) a. X wrote Y.
 b. X is the author of Y.

Reformulative paraphrasing occurs in a reformulation context when a rewording of a previously expressed content is added for discursive reasons, such as emphasis, correction, or clarification. Non-reformulative paraphrasing does not consider the role that paraphrasing plays in discourse. Reformulative paraphrase pairs have to be extracted from a single piece of discourse; non-reformulative paraphrase pairs can be extracted—each member of the pair on its own—from different discourse pieces. The reformulation in the third utterance in Example (8) gives an explanation in a language less technical than that in the first utterance; whereas Examples (9a) and (9b) are simply two alternative ways of expressing an authorship relation.

The strong discourse dependency of coreference explains the major role it plays in terms of cohesion. Being such a cohesive device, it follows that intra-document coreference, which takes place within a single discourse unit (or across a collection of documents linked by topic), is the most primary. Cross-document coreference, on the other hand, constitutes a task on its own in NLP but falls beyond the scope of linguistic coreference due to the lack of a common universe of discourse. The assumption behind cross-document coreference is that there is an underlying global discourse that enables various documents to be treated as a single macro-document.

Despite the differences, the discourse function of reformulative paraphrasing brings it close to coreference in the sense that they both contribute to the cohesion and development of discourse.

3. Mutual Benefits

Both paraphrase extraction and coreference resolution are complex tasks far from being solved at present, and we believe that there could be improvements in performance

if researchers on each side paid attention to the others. The similarities (i.e., relations of sameness, relations between NPs) allow for mutual collaboration, whereas the differences (i.e., focus on either meaning or reference) allow for resorting to either paraphrase or coreference to solve the other. In general, the greatest benefits come for cases in which either paraphrase or coreference are especially difficult to detect automatically. More specifically, we see direct mutual benefits when both phenomena occur either in the same expression or in neighboring expressions.

For pairs of linguistic expressions that show both relations, we can hypothesize paraphrasing relationships between NPs for which coreference is easier to detect. For instance, coreference between the two NPs in Example (10) is very likely given that they have the same head, head match being one of the most successful features in coreference resolution (Haghighi and Klein 2009). In contrast, deciding on paraphrase would be hard due to the difficulty of matching the modifiers of the two NPs.

- (10) a. The director of a multinational with huge profits.
 b. The director of a solvent company with headquarters in many countries.

In the opposite direction, we can hypothesize coreference links between NPs for which paraphrasing can be recognized with considerable ease (Example (11)). Light elements (e.g., *fact*), for instance, are normally taken into account in paraphrasing—but not in coreference resolution—as their addition or deletion does not involve a significant change in meaning.

- (11) a. The creation of a company.
 b. The fact of creating a company.

By neighboring expressions, we mean two parallel structures each containing a coreferent mention of the same entity next to a member of the same paraphrase pair. Note that the coreferent expressions in the following examples are printed in *italics* and the paraphrase units are printed in **bold**. If a resolution module identifies the coreferent pairs in Example (12), then these can function as two anchor points, *X* and *Y*, to infer that the text between them is paraphrastic: *X complained today before Y*, and *X is formulating the corresponding complaint to Y*.

- (12) a. *Argentina*_X **complained today before** *the British Government*_Y about the violation of the air space of this South American country.
 b. *This Chancellorship*_X **is formulating the corresponding complaint to** *the British Government*_Y for this violation of the Argentinian air space.

Some authors have already used coreference resolution in their paraphrasing systems in a similar way to the examples herein. Shinyama and Sekine (2003) benefit from the fact that a single event can be reported in more than one newspaper article in different ways, keeping certain kinds of NPs such as names, dates, and numbers unchanged. Thus, these can behave as anchor points for paraphrase extraction. Their system uses coreference resolution to find anchors which refer to the same entity.

Conversely, knowing that a stretch of text next to an NP paraphrases another stretch of text next to another NP helps to identify a coreference link between the two NPs, as shown by Example (13), where two diction verbs are easily detected as a paraphrase and thus their subjects can be hypothesized to corefer. If the paraphrase system

identifies the mapping between the indirect speech in Example (13a) and the direct speech in Example (13b), the coreference relation between the subjects is corroborated. Another difficult coreference link that can be detected with the help of paraphrasing is Example (14): If the predicates are recognized as paraphrases, then the subjects are likely to corefer.

- (13) a. *The trainer of the Cuban athlete Sotomayor* **said** that the world record holder is in a fit state to win the Games in Sydney.
 b. “The record holder is in a fit state to win the Olympic Games,” **explained** *De la Torre*.
- (14) a. *Police officers* **searched 11 stores in Barcelona**.
 b. *The investigators* **carried out 11 searches in stores in the center of Barcelona**.

Taking this idea one step further, new coreference resolution strategies can be developed with the aid of shallow paraphrasing techniques. A two-step process for coreference resolution might consist of hypothesizing first sentence-level paraphrases via *n*-gram or named-entity overlapping, aligning phrases that are (possible) paraphrases, and hypothesizing that they corefer. Second, a coreference module can act as a filter and provide a second classification. Such a procedure could be successful for the cases exemplified in Examples (12) to (14).

This strategy reverses the tacit assumption that coreference is solved before sentence-level paraphrasing. Meaning alone does not make it possible to state that the two pairs in Example (5b), repeated in Example (15), or the two pairs in Example (16) are paraphrases without first solving the coreference relations.

- (15) a. *The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona: “I had never beaten Barcelona.”
 b. *The Atlético de Madrid goalkeeper, A. Jiménez*, yesterday realized one of his dreams by defeating Barcelona, and said that *he* had never beaten Barcelona.
- (16) a. Secretary of State Colin Powell last week ruled out *a non-aggression treaty*.
 b. But Secretary of State Colin Powell brushed off *this possibility*.

However, cooperative work between paraphrasing and coreference is not always possible, and it is harder if neither of the two can be detected by means of widely used strategies. In other cases, cooperation can even be misleading. In Example (17), the two bold phrases are paraphrases, but their subjects do not corefer. The detection of words like *another* (Example (17b)) gives a key to help to prevent this kind of error.

- (17) a. A total of 26 Cuban citizens remain in the police station of the airport of Barajas **after requesting political asylum**.
 b. Another three Cubans **requested political asylum**.

On the basis of these various examples, we claim that a full understanding of both the similarities and disparities will enable fruitful collaboration between researchers working on paraphrasing and those working on coreference. Even more importantly,

our main claim is that such an understanding about the fundamental linguistic issues is a prerequisite for building paraphrase and coreference systems not lacking in linguistic rigor. In brief, we call for the return of linguistics to paraphrasing and coreference automatic applications, as well as to NLP in general, adhering to the call by Wintner (2009: 643), who cites examples that demonstrate “what computational linguistics can achieve when it is backed up and informed by linguistic theory” (page 643).

Acknowledgments

We are grateful to Eduard Hovy, M. Antònia Martí, Horacio Rodríguez, and Mariona Taulé for their helpful advice as experienced researchers. We would also like to express our gratitude to the three anonymous reviewers for their suggestions to improve this article.

This work was partly supported by FPU Grants AP2006-00994 and AP2008-02185 from the Spanish Ministry of Education, and Project TEXT-MESS 2.0 (TIN2009-13391-C04-04).

References

- Androutopoulos, Ion and Prodrimos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.
- Bhagat, Rahul. 2009. *Learning Paraphrases from Text*. Ph.D. thesis, University of Southern California, Los Angeles, CA.
- Callison-Burch, Chris. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh.
- Doddington, George, 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 the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon.
- Dolan, Bill, Chris Brockett, and Chris Quirk. 2005. README file included in the Microsoft Research Paraphrase Corpus, March, Redmond, WA.
- Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Sydney.
- Fuchs, Catherine. 1994. *Paraphrase et énonciation. Modélisation de la paraphrase langagière*. Ophrys, Paris.
- Fujita, Atsushi. 2005. *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Ph.D. thesis, Nara Institute of Science and Technology, Ikoma, Nara.
- Haghighi, Aria and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1152–1161, Singapore.
- Hirschman, Lynette and Nancy Chinchor. 1997. MUC-7 Coreference task definition—Version 3.0. In *Proceedings of the Message Understanding Conference-7 (MUC-7)*, Washington, DC.
- Hirst, Graeme J. 1981. *Anaphora in Natural Language Understanding: A Survey*. Springer-Verlag, Berlin.
- Karttunen, Lauri. 1976. Discourse referents. In J. McCawley, editor, *Syntax and Semantics*, volume 7. Academic Press, New York, pages 363–385.
- Madnani, Nitin and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36, pages 341–387.
- Milićević, Jasmina. 2007. *La paraphrase*. Peter Lang, Berne.
- Ng, Vincent. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, pages 1396–1411.
- Poesio, Massimo and Yannick Versley. 2009. Computational models for the interpretation of anaphora: A survey. Notes from the ACL-2009 Tutorial on State-of-the-art NLP Approaches to Coreference Resolution, Singapore.
- Recasens, Marta, Eduard Hovy, and M. Antònia Martí. 2010. A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 149–156, Valletta.
- Recasens, Marta and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated

- corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.
doi: 10.1007/s10579-009-9108-x.
- Shinyama, Yusuke and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the ACL 2nd International Workshop on Paraphrasing (IWP 2003)*, pages 65–71, Sapporo.
- van Deemter, Kees and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.
- Vila, Marta, Santiago González, M. Antònia Martí, Joaquim Llistnerri, and M. Jesús Machuca. 2010. CIIInt: A bilingual Spanish-Catalan spoken corpus of clinical interviews. *Procesamiento del Lenguaje Natural*, 45, 105–111.
- Wintner, Shuly. 2009. What science underlies Natural Language Engineering? *Computational Linguistics*, 35(4):641–644.

