Coreference Annotation: Whither?

Rodger Kibble and Kees van Deemter

Information Technology Research Institute University of Brighton Brighton BN2 4CJ U.K.

{Rodger.Kibble,Kees.van.Deemter }@itri.brighton.ac.uk

Abstract

The terms *coreference* and *anaphora* tend to be used inconsistently and interchangeably in much empirically-oriented work in NLP, and this threatens to lead to incoherent analyses of texts and arbitrary loss of information. This paper discusses the role of coreference annotation in Information Extraction, focussing on the coreference scheme defined for the MUC-7 evaluation exercise. We point out deficiencies in that scheme and make some suggestions towards a new annotation philosophy.

1. Introduction: coreference annotation for information extraction

Terms like coreference, anaphora and cospecification are used without clear definitions in much empiricallyoriented NLP/LE work such as information extraction, corpus analysis. We argue in this paper that this is not just a terminological issue, but that this lack of precision threatens to lead to incoherent analyses of texts and arbitrary loss of information. As a case study we show that these problems arise in one particularly well-known scheme, the MUC Coreference (CO) task (Hirschman and Chinchor, 1997a). A more detailed discussion of the problematic issues can be found in (van Deemter and Kibble, 1999). In this paper we concentrate on demonstrating how the problems in the MUC scheme arise from the requirements of the higherlevel IE tasks, and we argue that there is a need to split the CO task into more fine-grained tasks. We begin by attempting to clarify the notions of coreference and anaphora, using some textbook definitions (Trask, 1993) as a starting point. Notions of reference and coreference have had extensive discussion in semantics and philosophical logic for much of the last century (starting with (Frege, 1892/1960)), and we draw on this literature in (van Deemter and Kibble, 2000) in order to arrive at clear formulations of these notions. The present paper is more practically oriented, and considers how notions of coreference and anaphora can play a part in empirical applications such as information extraction and summarisation and how analyses of a text (by human annotators or systems) can be evaluated.

Finally, it is important to keep in mind that an annotation scheme consists of not only a set of tag definitions and instructions for annotators, but also a scoring procedure for measuring agreement between annotators or between a system and "truth" (i.e., a human annotation). An annotation scheme is only useful for practical applications if there is an agreed and objective method for evaluating and comparing different analyses of a text. So in section 0.0 we discuss some of the decisions that have to be made in choosing a scoring algorithm, and pay particular attention to the algorithms used by MUC participants and the constraints these algorithms impose on the types of relations which can be marked up.

1.1. Coreference and anaphora: some working definitions

Trask 1993 defines the key notions in this domain as follows:

- anaphor: An item with little or no intrinsic meaning or reference which takes its interpretation from another item in the same sentence or discourse, its antecedent. For example, in *I asked Lisa to check the proofs, and she did it*, the items *she* and *did it* are anaphors, taking their interpretations from their antecedents *Lisa* and *check the proofs*, respectively. (p15)
- **coreference:** The relation which obtains between two NPs (usually two NPs in a single sentence) both of which are interpreted as referring to the same extralinguistic entity. In linguistic representations, coreference is conventionally denoted by coindexing: *Lisa*_i said she_i would come. (p64-5)
- **reference:** The phenomenon by which some noun phrase in a particular utterance or sentence is associated with some entity in the real or conceptual world, its referent. (p232)

For present purposes, we will use these definitions, with only one modification: In accordance with present practise, we will widen the notion of anaphora to include cases where the anaphor does have intrinsic meaning, but not enough to refer uniquely on its own. For example, in

 BMW defends *its* decision to sell Rover which was costing *the German car company* one million pounds a day. (BBC News website 28 March, 2000)

the phrase *the German car company* narrows down the choice of referents to quite a small number of candidates.

Some clear distinctions are implicit in Trask's definitions:

• Coreference requires reference whereas anaphora does not. So, one can in principle have anaphora without coreference. Examples of non-referring NPs arguably include bound anaphors and NPs in hypothetical or negative contexts (van Deemter and Kibble 1999, 2000).

- NPs can be coreferential without being anaphoric, if each NP refers independently without a dependence on the other. This happens most clearly when two proper names or descriptions are used on two different occasions (or in parts of the corpus that are far apart), so that there is no possibility of one influencing the interpretation of the other.
- Coreference is an equivalence relation (i.e., it is reflexive, transitive and symmatrical) whereas anaphora is not. For example if NP₁ is anaphoric to NP₂ the reverse does not generally hold.

1.2. Evaluating coreference annotations

Coreference annotation proceeds by first assigning a unique identifier to each of the relevant NPs. For each NP (say x) of these, the annotator may specify an identifier (say n) as the value of the REF feature of x if and only if she decides that x corefers with the NP that has n as its identifier. On the face of it, this procedure makes coreference a non-symmetrical, nontransitive, and nonreflexive relation but, in fact, the coreference relation is construed as the reflexive, symmetrical and transitive closure of the relation induced by the REF feature. (See also (Hirschman et al., 1997) for confirmation of this interpretation.) Thus, the only thing that counts in an annotation is the equivalence classes (induced by the relation 'NP₁ corefers with NP₂') that it gives rise to. Consider, for example, the sentence

(2.) General Pinochet was coming round from an operation for a lumbar hernia in a London hospital when he^1 was given the bad news that he^2 was being detained.

One annotator might mark up coreference pairs $\langle General Pinochet, he^1 \rangle$, $\langle he^1, he^2 \rangle$, and the second $\langle General Pinochet, he^1 \rangle$, $\langle General Pinochet, he^2 \rangle$. Both annotations lead to the same coreference relation. One way to see this is by checking that the transitive, symmetrical and reflexive closures of the two relations are equal.

This 'equivalence class' interpretation of coreference annotation is also reflected by the way in which annotations are compared. The degree to which two coreference annotations agree (each of which may be done by hand or generated by computer) can be measured in different ways. Each of these, however, is based on the assumption that the relation in question is an equivalence relation. This is especially evident in the B-CUBED scoring algorithm (Baldwin et al., 1997), which seeks to go beyond traditional measures of precision and recall. This is best explained comparing two fictitious annotations. Consider the following text:

Ocean Drilling & Exploration Co.(1) will sell its(1) contract-drilling business(2), and took a \$50.9 million loss from discontinued operations in the third quarter(3) because of the planned sale.

The New Orleans oil and gas exploration and diving operations company(1) added that it(1) doesn't expect any further adverse financial impact from the restructuring.

In the third quarter(3), the company(1), which is 61%-owned by Murphy Oil Corp. of Arkansas, had a net loss(4) of \$46.9 million(4), or 91 cents a share(4).

It has long been rumored that Ocean Drilling(1) would sell the unit(2) to concentrate on its(1) core oil and gas business.

Suppose System A interprets The New Orleans oil and gas exploration and diving operations company as the name of a company and doesn't mark it as identical to Ocean Drilling. Compare this with System B which has a naive approach to resolving pronouns, and always picks the most recent entity as an antecedent: this works for two of the three pronouns but would identify the third one with the unit rather than Ocean Drilling. Both systems have the same recall (measured by the number of links which would minimally have to be added to correct the annotation) error, and in addition System B has a small precision error (measured by the number of links that would have to be removed). Baldwin, however, reflecting on situations of this kind, argued convincingly that the error made by a system behaving like A is the more damaging, because what counts is how many coreference links can be *inferred* from each annotation by taking into account that coreference is an equivalence relation (in particular, a transitive relation). To reflect this, (Baldwin et al., 1997) propose a new scoring algorithm B-CUBED in which scores are weighted according to the size of equivalence classes in the response and answer keys. As a result, System A would be penalised for breaking up a long coreference chain into two shorter ones.

Although evaluation algorithms share the assumption that coreference is an equivalence relation (see also (Vilain et al., 1996)), we will show that the IDENT relation in MUC is actually a *family* of coreference/anaphora relations some of which are *asymmetric* and so do not result in an equivalence relation.¹

2. The MUC-7 Coreference task

The Message Understanding Conferences (MUC) are a high-profile series of evaluation exercises in which institutions are invited to compete on a few clearly defined IE tasks, using datasets and scoring algorithms provided by the MUC developers. The Coreference task was introduced in the MUC-6 exercise and a refined version was included in the latest evaluation round, MUC-7 (Hirschman and Chinchor, 1997b). The criteria for the Coreference task are listed as follows (Hirschman and Chinchor, 1997a):

- 1. Support for the MUC IE tasks
- 2. Good inter-annotator agreement ca. 95%

¹Note that this problem cannot be simply tackled by focusing on the reflexive, symmetrical and transitive closure of the different relations involved (such as anaphora) since even the combination of two equivalence relations is not always an equivalence relation (van Deemter and Kibble 1999). For example, if NP₁ and NP₂ stand in an *anaphoric* relationship and NP₂ and NP₃ *corefer* then it is possible that NP₁ and NP₃ neither corefer nor stand in an anaphoric relationship.

- 3. Quick/cheap markup procedures
- 4. Create a resource for the research community

Additional goals:

- "priority on preserving reasonable semantics for the equivalence classes."
- "judgments should be based on the intelligent reader's knowledge of the world resulting from his or her best understanding of the text...not...on a linguistic theory of how NPs are resolved..."

There is admitted to be some potential conflict between these goals, and we will see that there is actually a tension within item (1), between the different IE tasks themselves.

2.1. MUC IE tasks

The tasks which are most directly concerned with coreference resolution are Named Entity, Template Element and Scenario Template. The various tasks are briefly summarised below (see (Hirschman and Chinchor, 1997b) for full specifications).

Named Entity (NE) task

The task is to identify and categorise all instances in a text of the following kinds of expression: *names* of persons, organisations or locations; *times*, i.e. dates and times of day and *quantities* - money values, percentages.

Coreference (CO) task

This task involves marking the IDENT relation between NPs which are judged to be "coreferential". CO differs from the other tasks in that the end product is not a template but a set of annotations within the text. Thus, CO does not produce information which will be directly presented to an end user, but provides raw materials for the other tasks to construct database records by filling slots in templates of different kinds. In other words, coreference analysis is not an end in itself but mediates between the NE, TE and ST tasks as described below.

Attributes: Template Element (TE) task

The TE task collects attributes of entities mentioned in a text, including both those identified by NE and entities which are only referred to by a description. This is done by extracting descriptors from NPs which are marked as coreferential with the entity in question by the CO task. In order to include all available descriptors, coreference is stipulated to cover predicative and appositive NPs (e.g., *General Wesley Clark, Nato's supreme commander*..) although other analysts such as Passoneau (1997) do not class these as coreferential.

Facts: Template Relation (TR) task

The TR task marks relationships between Template elements. The relations included in MUC-7 are limited to LOCATION_OF, EMPLOYEE_OF, PRODUCT_OF. This has no direct connection with CO and is mentioned here only for completeness.

Events: Scenario Template (ST) task

The ST task identifies events in which entities participate. The scenarios involved in MUC-7 consist of missile launch reports, but other typical applications include management succession reports as illustrated in example (6). The function of CO in support of this task is to enable all events in which an entity participates to be identified, by defining equivalence classes of all NPs which refer to each discourse entity mentioned in a text.

We see already that CO involves two distinct uses of NPs: referential uses identifying arguments of predicates, and descriptive uses which convey information about an entity other than the identity of its referent.

Example (5) illustrates these distinctions:

- (3.) General Pinochet (1) left the U.K. yesterday. The former military dictator (2) had been found unfit to stand trial. Pinochet (3) seized power in a military coup in 1973 and the following year he (4) declared himself (5) president (6).
- **Referential** mentions supporting ST task: (1), (2), (3), (4), (5)

Descriptive mentions supporting TE task: (2), (6)

Both of these are classed as IDENT in the MUC TD. We demonstrate in the following sections how this threatens to lead to incoherence and loss of information in coreference annotations.

2.2. Grounding: extensional and intensional descriptors

Section 6.4 of the TD tells annotators that 'Two markables should be recorded as coreferential if the text asserts them to be coreferential at *any time*'. Accordingly, in example (6):

(4.) Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents. Sudsy Soaps named Eliza Dolittle as sales director effective last week.

annotators are asked to mark the pairs (1) Henry Higgins and (2) sales director of Sudsy Soaps, and (1) Henry Higgins and (3) president of Dreamy Detergents as respectively standing in the IDENT relation. But if we treat this relation as coreference, this implies by transitivity that the sales director of Sudsy Soaps is the president of Dreamy Detergents, which is not what the text asserts. And if the names Henry Higgins and Eliza Dolittle are analysed as co-referring with sales director of Sudsy Soaps and sales director respectively, there is a danger that these two descriptions may be interpreted as synonymous, leading to a collapsing coreference chain where Henry Higgins is asserted to be identical with Eliza Dolittle. This problem was noted following the experience of MUC-6, when the guidelines were interpreted by some annotators as requiring mentions of the same position in the same company to be marked as coreferential (Hirschman et al., 1997) and this did indeed result in "collapsing coreference".

In fact the TD is designed to avoid these predictions, and the idea that IDENT is strict coreference is tacitly dropped. Hirschman et al (1997) introduce the distinction between *extensional* mentions (reference to individuals by name) and *intensional* mentions (reference by description, e.g. job title or military rank).

Intensional mentions, such as *CEO*, are *grounded* by association with extensional mentions (e.g., *Mr Donner*) which prevents the collapse of coreference chains.

(Hirschman et al., 1997)

So the above example has three coreference chains, each grounded in a different existential description:

- [Henry Higgins, sales director for Sudsy Soaps, president of Dreamy Detergents]
- [Eliza Dolittle, sales director]

[Sudsy Soaps, Sudsy Soaps]

This analysis has some unfortunate consequences. The IDENT relation turns out to be considerably weaker than strict coreference: members of "coreference chain" only *potentially* corefer. Each member describes the grounding instance *at some time*, but an arbitrary pair may or may not "corefer" (describe the same individual) at any particular time. This results in a loss of information in various ways:

- IDENT is so weakly specified that genuine coreference links can't be inferred
- Annotators' decisions on whether an NP is "extensional" or "intensional" are not recorded: everything is marked as IDENT, making it harder to analyse disagreement.
- Contradictions can't be identified: if another source reports Higgins became the president of Divine Deodorants, the result will simply be an extended "coreference chain" as follows:

[Henry Higgins, sales director for Sudsy Soaps, president of Dreamy Detergents, president of Divine Deodorants]

In order to come up with a remedy for this state of affairs, we need to recognise that IDENT covers at least two distinct tasks: coreference proper and marking up elements of a Composite Description (CD), supporting ST and TE respectively. To illustrate this: example (7) includes an appositive NP *General Wesley Clark, Nato's supreme commander*, and analysts have differed over whether the two component NPs should be marked as coreferential. The MUC scheme requires that they are marked as coreferring, though (Passoneau, 1997) for example argues that they should not be. We would argue that there is nothing to be gained from marking items (1) and (2) as *coreferential* since they occupy the same argument role, but that they both form part of a CD made up of attributes of the individual identified as a Named Entity *General Wesley Clark*.

(5.) General Wesley Clark (1), Nato's supreme commander (2), immediately ordered 500 British and French paratroopers (3) to be put on standby to occupy the airport (4).

2.3. Functional values

The way intensionality is handled in the MUC Coreference TD leads to further complications when we come to consider numerical values. For example, in Section 1.3 of the TD, concerning the implications of 'change over time', where the example *the stock price fell from \$4.02 to \$3.85* is discussed, annotators are asked to consider *the stock price* as standing in the IDENT relation with \$3.85 but not with \$4.02, because \$3.85 is 'the more recent value'. Quite reasonably, \$4.02 is not considered to stand in the IDENT relation with *the stock price* because transitivity would lead to the conclusion that \$4.02 and \$3.85 are equal. But, what if the price continues to fall?

- (6.) a. *The stock price* fell from \$4.02 to \$3.85;
 - b. Later that day, it fell to an even lower value, at \$3.82.

Does the annotator have to go back to (a), deciding that \$3.82 is an even more recent value and *the stock price* does not stand in the IDENT relation with \$3.85 after all? This rather seems to contradict the instruction in Section 6.4 that IDENT holds between two items "if the text asserts them to be coreferential at *any time*". Clearly, the issue of dealing with change over time and modality (van Deemter and Kibble 2000) is in need of rethinking.

Another, subtler problem is also worth noting, which has to do with the way in which annotators are asked to treat numbers as grounding instances. Consider a sentence like the following:

(7.) The UK satellite television broadcaster said *its sub*scriber base grew ... to 5.35 million.

Suppose the population of Scotland is also 5.35 million. If annotators understand the Named Entity 5.35 million as grounding both *the population of Scotland* and *the subscriber base of* ... then it follows that the subscriber base is the same as the population of Scotland. This anomaly is caused by the fact that there is a hidden *cardinality* function that is not expressed explicitly by the text. Note that it *would* be correct to say that

the cardinality of (the subscriber base of (b)) = the cardinality of (the population of (s)).

This shows that treating numerical values as grounding instances can be a hazardous affair even apart from issues arising from change over time. One might even question whether treating numerical quantities as instantiations of descriptive terms is appropriate to the requirements of IE: a collection of quantity terms which all happen to have the value e.g. \$4.02 is not likely to be of much interest to users of IE systems. The task as formulated will not support a search on e.g. all recorded values of a particular stock price, just as it does not allow for queries about "all people who have been president of a particular company" (Hirschman and Chinchor, 1997a, p. 2). So from the point of view of practicality as well as theoretical consistency, there appears to be a need for a new strategy for dealing with values of functional or time-dependent descriptions. We briefly return to this point in Section 3.

2.4. Bound anaphora

Annotators are also instructed to mark an IDENT link between a bound anaphor and the NP which binds it. It is not clear why this is included in the task: the authors admit that "one may argue that such elements are not coreferential in the usual sense" (Hirschman and Chinchor, 1997a, p. 10), and marking these relations does not directly support any of the IE tasks. That is, information from quantified noun phrases does not provide information about individuals unless the system is capable of doing some reasoning, and information about individuals can only be weakly inferred in the case of quantifiers such as *most, many*.

It is straightforward to demonstrate via a substitution test that bound anaphors do not corefer with their antecedents. Consider, for example, quantifying NPs such as 'Every TV network' (or, even more problematic, 'Most computational linguists' (Hirschman and Chinchor, 1997a)). If 'Every TV network' refers at all, then presumably it refers to the set of all TV networks (relevant to a certain domain). The TD, however, asks annotators to let 'Every TV network' corefer with its in 'Every TV network reported its profits'. But this implies that, in extensional contexts, the two NPs are interchangeable. In other words, (10a) below must mean the same as (10b), which is clearly false.

- (8.) a. Every TV network reported its profits.
 - b. Every TV network reported every TV network's profits.

3. Discussion

We will conclude by summarising the problems we have identified with the MUC-7 coreference scheme and by making some practical proposals for a more coherent but perhaps less ambitious annotation philosophy (if not yet a new annotation scheme).

3.1. Summary of shortcomings

- The Coreference relation as specified by MUC does not result in an equivalence class - IDENT is not in general transitive or symmetric. There is a tension between a scoring regime which assumes equivalence classes and IE requirements which bring in asymmetric relations, e.g. extensional/intensional descriptors, "grounding instance" etc. This leads to some doubt about how informative the resulting scores can be.
- 2. "Coreference" as specified by MUC includes nonreferring NPs such as quantifiers and bound anaphors.
- 3. There is a loss of information as to why decisions were taken, although annotators are expected to recognise phenomena such as bound anaphora and relations between functions such as temperatures, prices and their numerical values. All links are simply flagged as IDENT, making it harder to analyse disagreements.

4. Treatment of functional values results in arbitrary loss of information (stock price, temperature). This is inconsistent with the treatment of time-dependent roles for persons.

3.2. Summary of proposals

In (van Deemter and Kibble, 1999; van Deemter and Kibble, 2000) we proposed that the CO task be re-oriented to cover genuine coreference only, possibly including clausal as well as nominal reference. We postulated a more sophisticated strategy which would require annotators to distinguish between "individual concepts" and "individuals" but wondered whether this might be "asking too much".

In fact the MUC-7 TD already requires annotators to make a similar distinction between "intensional" and "extensional" descriptors in deciding whether to mark a coreference link. However these decisions are not recorded in the annotation, making it harder to analyse disagreements. The MUC IE tasks, in particular the TE task, depend on retrieving the content of descriptive NPs - this is the highest priority criterion for the CO task. We list some prerequisites below for the design of a new annotation scheme, distinguishing between the annotation tasks which are currently subsumed under CO and the template tasks for which it provides the input data.

- Formulate semantically coherent definitions of reference and coreference without losing sight of the requirements of data-oriented applications such as IE. (van Deemter and Kibble, 2000) discuss two possible options: identifying a *referring* NP with a *semantically definite* NP in the sense of (Barwise and Cooper, 1981), or a more liberal approach which treats a quantified NP as a referring expression which picks out a set of which the sentence as a whole is true: e.g. the subject of '*Most computational linguists use a parser*' would refers to the set of those computational linguists *who use a parser* (cf. (Kamp and Reyle, 1993)).
- Recognise that the CO annotation task as specified in the MUC task has two distinct goals: to identify elements of *composite descriptions* for the purposes of the TE task, and to identify equivalence classes of extensional references in support of the ST task. These tasks need to be separated into two distinct annotation tasks: coreference proper, and a new task CD which links descriptive NPs (president of Sunlight Soaps) with NPs which identify discourse entities directly (e.g., Henry Higgins). These links will often overlap with coreference links since as we saw earlier NPs may simultaneously refer and have descriptive content. However, we would want to exclude from coreference proper predicative NPs which appear as arguments of copular verbs or in apposition, for instance. Example (9) illustrates a possible annotation using a new tag DESC to relate descriptive and referring NPs:
 - (9.) <COREF ID = 1>Henry Higgins</ID> has become <DESC SUBJ = 1>president of Dreamy

Detergents</DESC>. <COREF ID = 2 REF = 1> He </COREF> has resigned from Sudsy Soaps.

- This extension of the annotation tasks will need to be reflected in a reorientation of the template tasks. Currently TE is rather weakly specified, with one slot for NAME derived from the Named Entity task and one for DESCRIPTOR, which must not have more than one value. Recall that NAME may be a numerical value and DESCRIPTOR a term such as stock price. Alternatively, NAME could identify a person with DE-SCRIPTOR being a job title, army rank etc. To support efficient processing of queries such as What was the highest price of Microsoft stocks on March 31st?, it would be more useful to fill in a Function-Value template using the links annotated by CD, constructing database records with a fixed header portion identifying the TYPE and (optionally) NAME of the functional expression and a variable-length list containing all recorded VALUES.
- The above point raises the question of how functionvalue relations are to be characterised in a way which is semantically respectable and can be straightforwardly explained to annotators. A strategy which would be consistent with (Dowty et al., 1981)) (based on an analysis in (Frege, 1892/1960)) would be to say that *The stock price* refers to a Montague-type individual concept, that is, a function from times to numbers. It would have followed that *The stock price* does not corefer with either \$4.02 or \$3.85 and no problem would have arisen. Analogously, *president of Dreamy Detergents*, in the context cited above, would denote an individual concept rather than an individual. (These issues are discussed in more detail in (van Deemter and Kibble, 2000).)
- Finally, if it is considered necessary for IE to annotate bound anaphora, this task will need its own TD, though as we pointed out in section 2.4. there is no direct link between this task and the IE tasks.
- As we stressed in Section 1.2., an annotation scheme only has practical value if there is an agreed, objective procedure for comparing and evaluating different annotations of the same text. The scoring algorithms used in the MUC exercises have assumed that IDENT is an equivalence relation, and we have seen that IDENT is actually a family of relations some of which are asymettric. So there is a need to develop a new scoring regime as a prerequisite for any further refinements of the annotation scheme itself.

Acknowledgements

We are grateful to Lynette Hirschman and Breck Baldwin for their very constructive responses to a presentation on the topic of this paper (van Deemter and Kibble, 1999). Rodger Kibble's participation in this research was funded by the UK EPSRC as part of the GNOME (GR/L51126) and RAGS (GR/L77102) projects.

4. References

- Baldwin, Breck et al., 1997. Description of the UPenn Camp system as used for coreference. In *Procs. of 7th Message Understanding Conference*. See www.muc.saic.com.
- Barwise, Jon and Robin Cooper, 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4.
- Dowty, David, Robert Wall, and Stanley Peters, 1981. Introduction to Montague Semantics. Dordrecht: Kluwer.
- Frege, Gottlob, 1892/1960. On sense and reference. In Peter Geach and M Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell.
- Hirschman, Lynette and Nancy Chinchor, 1997a. Muc-7 coreference task definition. In *MUC-7 Proceedings*. Science Applications International Corporation. See www.muc.saic.com.
- Hirschman, Lynette and Nancy Chinchor (eds.), 1997b. *MUC-7 Proceedings*. Science Applications International Corporation. See www.muc.saic.com.
- Hirschman, Lynette, P Robinson, J Burger, and M Vilain, 1997. Automating coreference: The role of annotated training data. In *Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing..*
- Kamp, Hans and Uwe Reyle, 1993. From Discourse to Logic. Dordrecht: Kluwer.
- Passoneau, Rebecca, 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.
- Trask, R. L., 1993. A Dictionary of Grammatical Terms in Linguistics. London and New York: Routledge.
- van Deemter, Kees and Rodger Kibble, 1999. What is coreference and what should coreference annotation be? In Amit Bagga, Breck Baldwin, and S Shelton (eds.), *Procs. of ACL workshop on Coreference and Its Applications*. Maryland.
- van Deemter, Kees and Rodger Kibble, 2000. On coreferring; coreference in MUC and related annotation schemes. Submitted to *Computational Linguistics*.
- Vilain, Marc et al., 1996. A model-theoretic coreference scoring scheme. In *Procs. of 6th Message Understanding Conference*. See www.muc.saic.com.