MUC-7 Coreference Task Definition

Version 3.0

13 July 1997

Source: Lynette Hirschman (lynette@MITRE.org)
Keeper: Nancy Chinchor (chinchor@gso.saic.com)

1. **INTRODUCTION**

1.1 Rationale for the Coreference Task Definition

The task definition has been constructed to capture one level of information, coreferring expressions, in the context of the Message Understanding Conference (MUC) information extraction tasks. The coreference "layer" links together multiple expressions designating a given entity; for now, only nouns are linked -- relations involving verbs are ignored. The coreference layer functions to collect together all mentions of a given entity, including those tagged in the Named Entity task. We can look at coreference annotation as potentially supporting a kind of hyperlinked version of the text, where the links connect the mentions of a given entity.

In the context of MUC, the coreference layer provides input to the template element task, where each named entity is represented as a single template which collects information about that element from the multiple mentions in the text. Similarly, coreference provides input to the scenario template, especially where that task requires filling the template with entries other than named entities.

1.2 Criteria for the Task Definition

There are four criteria for the current task definition, which often push in different directions:

1) Support for the MUC information extraction tasks;

2) Ability to achieve good (ca. 95%) interannotator agreement;

3) Ability to mark text up quickly (and therefore, cheaply);

4) Desire to create a corpus for research on coreference and discourse phenomena, independent of the MUC extraction task.

These criteria are listed in order of priority -- and the decisions in the task definition have been made with these priorities in mind. In particular, we have tried for an extensive but not exhaustive coverage of coreference phenomena.

Based on experience in defining annotation schema for other phenomena, it is more important to preserve high inter-annotator agreement than to capture every possible phenomenon that could fall under the heading of "coreference". For this reason, the annotation scheme covers only the "IDENTITY" (or IDENT) relation for noun phrases; it does not include coreference among clauses, nor does it cover other kinds of coreference relations (set/subset, part/whole, etc.) If this task follows the same path as other annotation schemes, it should be possible to expand the task definition as consensus emerges (for both annotation and scoring) on these other phenomena.

However, there are trade-offs. In the case of the first priority, supporting MUC information extraction, it may not always be possible to align the coreference task to support the other tasks perfectly. For example, if the Scenario Template requires resolution of type relations, such as all people who have been president of a particular company, the coreference task definition may not support this directly (see section 4.2 and section 6.4 for a further discussion of this issue) .

1.3 Coreference Chains and Scoring Considerations

The coreference annotation scheme described below is focused on the IDENTITY (IDENT) relation. This relation has important semantic properties that play a central role in defining the IDENT scoring mechanism.

The IDENT relation is symmetrical (if A is IDENT to B, then B is IDENT to A), and it is transitive (if A is IDENT to B and B is IDENT to C, then A is IDENT to C, and C is IDENT to A). These properties induce a set of EQUIVALENCE CLASSES among the marked elements, where each element participates in exactly one equivalence class, and all elements in an equivalence class are coreferring.

The IDENT relationship is NOT directional. Note that this is different from a part-whole or set-subset relation, which are ASYMMETRICAL and thus require a different scoring algorithm which interacts in complex ways with the IDENT relation; this is one reason why we will not attempt, in this round of revisions, to tackle these relations.

The nature of the IDENT relation and its associated coreference equivalence classes pose a problem where an expression may be coreferential with either of two NPs, because of conjunction, or because of type/instance ambiguity or in expressions of change over time. For example, in the sentence, "the stock price fell from $4.02 to $3.85", the stock price at one time is coreferential with $4.02, and at a later time with $3.85. However, if we make both $4.02 and $3.85 coreferential with stock price, we get "collapsing coreference chains" -- that is, we end up with *stock price*, *$4.02* and *$3.85* all in the same equivalence class -- which is counter-intuitive, and would prevent the IDENT relation from supporting, e.g., the Template Element task. This issue is discussed in some detail, with a number of examples, in section 4.2 and section 6.4.

In keeping with having the coreference task support other information extraction tasks, we propose to place highest priority on preserving reasonable semantics for the equivalence classes. This means that two values (or instances) that are clearly distinct should NOT be allowed to merge into an equivalence class, even if this means not being able to mark all of the function/value or type/instance relations we might want to mark. Thus, in the example above, we would mark *stock price* and the more recent value *$3.85* as coreferential, and leave *$4.02* in its own equivalence class, not marked coreferential with *stock price*. This means that the mark-up fails to capture some information (that *$4.02* is also a value, at an earlier time, of *stock price*), but this seems like a reasonable price to pay for preserving the semantics of the coreference equivalence classes.

These issues are related to our decision (for now) to mark only IDENT relations, with no distinction between types, functions, and instances. The advantage of this solution is that it leaves the conventions and the scoring mechanism intact and does not require additional mark-up for new kinds of coreference relations. At a future time, if we wish to distinguish type and function coreference from the IDENT relations, it should be possible to mark these relations more completely.

1.4 Future Directions

Given these limitations, we can see that the coreference annotation scheme can evolve in several directions. It would be useful to expand the annotation scheme for future MUCs to include:

1) coreference to cover clause (verbal) level relations

2) a method for handling discontinuous elements, including conjoined elements

3) a distinction between function/type coreference and instance coreference, which has caused some problems with the unintended merging of coreference chains

4) set/subset coreference, part/whole and other kinds of coreference

2. GENERAL NOTATION

2.1 SGML Tagging

The annotation for coreference is SGML tagging within the text stream. Referring expressions and their antecedents are tagged as follows:

<COREF ID="100">Lawson Mardon Group Ltd.</COREF> said <COREF ID="101" TYPE="IDENT" REF="100">it</COREF> ...

The basic annotation contains the information to establish some type of link between an explicitly marked pair of noun phrases. In the above example, the pronoun "it" is tagged as referring to the same entity as the phrase, "Lawson Mardon Group Ltd."

There is one markup per string. Other links can be inferred from the explicit links. We assume that the coreference relation is symmetric and transitive, so if phrase A is marked as coreferential with B (indicated by a REF pointer from A to B), we can infer that B is coreferential with A; if A is coreferential with B, and B is coreferential with C, we can infer that A is coreferential with C.

2.2 The "TYPE" Attribute

The purpose of the TYPE attribute is to indicate the relationship between the anaphor and the antecedent. At present only one such relationship, "IDENT" (for identity), is being annotated.

2.3 The "ID" and "REF" Attributes

The ID and REF attributes are used to indicate that there is a coreference link between two strings. The ID is arbitrarily but uniquely assigned to the string during markup. The REF uses that ID to indicate the coreference link.

2.4 The "MIN" Attribute

The MIN attribute is used in the answer key ("key") to indicate the minimum string that the system under evaluation must include in the COREF tag in order to receive full credit for its output ("response"). So, in the next example, if the system response had omitted "of Surrey, England" from the COREF tag, the response would nonetheless receive full credit because it identified the minimum string.

<COREF ID="100" MIN="Haden MacLellan PLC">Haden MacLellan PLC of Surrey, England</COREF>

... <COREF ID="101" TYPE="IDENT" REF="100">Haden MacLellan</COREF>

Any response which includes the MIN string and does not include any tokens beyond those enclosed in the <COREF>...</COREF> tags is valid. The MIN string will in general be the HEAD of the phrase; see section 5 for a full discussion of this issue. Note that only the annotation KEY distinguishes between the maximal string and the MIN string; the response key does not have a MIN attribute.

2.5 The "STATUS" Attribute

The STATUS ("status") attribute is used in the answer key when the markup is optional. The only value for this attribute is OPT ("optional").[1] The evaluation software will not score a string that is marked OPT in the key unless the response has markup on that string. A potential example is given below. (It is marked OPT because a reader may not be certain that "Livingston Street" refers to the Board of Education.) Note that the optionality is marked only for the anaphor.

[1]At the Feb. 96 meeting of the Coreference and Ellipsis working group, the suggestion was made to distinguish markups that are optional because of textual ambiguity from markups that are optional because of unclear or missing markup guidelines. Although this seems a workable suggesion, a little experimentation may be advisable before implementation.

Our <COREF ID="102" MIN="Board of Education">Board of Education</COREF> budget is just too high, the Mayor said. <COREF ID="103" STATUS="OPT" TYPE="IDENT" REF="102">Livingston Street</COREF> has lost control.

3. WHAT PART OF THE TEXT TO ANNOTATE

Coreference markup should be made on the body of the text and on corpus-specific portions of the header. The SGML tags that are used to identify the body and the various portions of the header may vary from one corpus to another.

3.1 Specific Guidance for MUC-7 Corpus

The annotation of coreference is to be performed within the text delimited by the SLUG, DATE, NWORDS, PREAMBLE, TEXT, and TRAILER tags.

3.2 Specific Guidance for Speech Transcriptions

If the transcript contains disfluencies or verbal erasures, the "erased" portion should not be annotated for coreference; this means that it will be helpful to have the input text annotated for disfluencies before beginning coreference annotation, so that there is agreement on what is "verbally deleted" and what is part of the final output.

4. WHAT THINGS TO ANNOTATE

4.1 Markables

The coreference relation will be marked between elements of the following categories: NOUNS, NOUN PHRASES, and PRONOUNS. Elements of these categories are MARKABLES. PRONOUNS include both personal and demonstrative pronouns, and with respect to personal pronouns, all cases, including the possessive. Dates ("January 23"), currency expressions ("$1.2 billion"), and percentages ("17%") are considered noun phrases.

A noun phrase is markable whether it is the object of an assertion, a negation, or a question. Thus, "a machete" is markable in all of the following examples:

I have a machete.

I don't have a machete.

Do you have a machete?

Note in particular that the initial introduction of an object into the discourse may often occur as an indefinite noun phrase ("Do you have a machete?" or "I saw *a truck*; *it* turned the corner..."). Also note that just because an element is "markable", it does not follow that there are later references to it -- that is, it may or may not participate in coreference. That may even be true for pronouns -- section 4.5 for further discussion.

Interrogative "wh-" noun phrases are NOT markables, e.g. "Which engine" and "Who" in the following queries:

Which engine would you like to use?

Who is your boss?

The relation is marked only between pairs of elements both of which are markables. This means that some markables that look anaphoric will not be coded, including pronouns, demonstratives, and definite NPs whose antecedent is a clause rather than a markable. For example, in

Program trading is "a racket," complains Edward Egnuss, a White Plains, N.Y., investor and electronics sales executive, "and *it's not to the benefit of the small investor*, *that*'s for sure."

Though "that" is related to "it's not to the benefit of the small investor", the latter is not markable, so no antecedent is annotated for "that".

4.2 Terminology for Mark-Up

It is useful to define some terms to support a discussion of the difficult coreference issues. This section defines the terms "extensional descriptor", "intensional descriptor" and "grounding instance".

An extensional descriptor is an enumeration of the member(s) of a set by (unique) names. In the context of the coreference task definition, this amounts to the use of proper names, e.g., *Jane Z. Smith*, *Chrysler Corporation*, or numerical values (The stock price was *$4.02*).

An intensional description is a predicate that is true of an entity or set of entities -- that it, it characterizes or defines the members of the set: "the prime numbers", "the president of Chrysler Corporation". Any non-concrete

common noun, taken on its own, is an intensional description: it functions at the "type" level ("president", "problem") or, if it takes a quantifiable value, at the "function" level ("rate", "temperature"). Intensional descriptions are useful for sets which have no finite extension ("the set of odd numbers"), or in cases where we don't know the extension ("the gene sequence responsible for encoding the immune response"). They can also be used to refer to instances of the type (*my Ford*... *the car*), or to values of the function (the [per share value of [$4.02]] ... [The stock price]).

The grounding instance in a coreference chain is the first extensional description in the chain (most often, the first element in the chain). This terminology is useful in the discussion about function-value relations, time-dependent entities and bare nominals. Thus in the sequence

Henry Higgins, who was formerly sales director for Sudsy Soaps, became president of Dreamy Detergents

we have a sequence consisting of the extensional description *Henry Higgins* (which is the grounding instance), together with two intensional descriptions, *sales director for Sudsy Soaps* and *president of Dreamy Detergents*. In addition, there are two other extensional descriptions, *Sudsy Soaps*, and *Dreamy Detergents*.

In the sentence

The temperature rose to 90 degrees before dropping to 70 degrees.

we have a function, "temperature", which takes on a value ("90 degrees") at one point in time, and at a later point in time, a second value, "70 degrees". Because there is only one occurrence of the noun phrase "the temperature", we have a problem marking the coreferring expressions. "The temperature" is a function expression, and is grounded first by "90 degrees"; an implicit second occurrence is coreferential with "70 degrees". However, if we mark all of these as coreferential, we find that "90 degrees" is IDENT with "70 degrees", which is clearly wrong. What we want to say is:

90 degrees instantiates "The temperature" at time t1; 70 degrees instantiates "The temperature" at time t2; And these are both of type "temperature" (but not IDENT).

In section 1.3, we proposed conventions that prevent the collapsing of coreference chains, at the expense of losing some type coreference. Given that our mark-up conventions are already incomplete (we don't mark verb coreference), this seems like a small price to pay for making the chains we do mark useful in other information extraction tasks, e.g., the Template Element task. We provide marked-up examples of this in section 6.4.

4.3 Names and Other Named Entities

Names and other Named Entities (as defined in the MUC-6 document titled "Named Entity Task Definition" -- dates, times, currency amounts, and percentages) are all markables. A substring of a Named Entity, however, is not a markable. Thus in

London ... *London*-based ...

the two instances of London are to be marked coreferential; in

Reuters Holding PLC ... *Reuters* announced that

"Reuters Holding PLC" and "Reuters" are to be marked coreferential. But in

Equitable of Iowa Cos. ... located in Iowa.

the two instances of "Iowa" are NOT to be marked as coreferential since the first is not a markable: it is a substring of a Named Entity. In addition to names as defined for the Named Entity task, other identifiers that are, in the opinion of the annotator, clearly not decomposable should be treated as atomic as well, e.g., "Widener Library" and "E two" in

<COREF ID="0" MIN="building">the large strange-looking building, which is <COREF ID="1" TYPE="IDENT">Widener Library</COREF></COREF>

and in

okay then I'll take <COREF ID="0" MIN="E two">engine E two</COREF> ... so uh the plan is to take <COREF ID="1" TYPE="IDENT" REF="0" MIN="E two">engine E two</COREF> ...

Date expressions recognized by the Named Entity task are also treated as atomic; components of a date are not separate markables. Thus, in

In a report issued January 5, 1995, the program manager said that there would be no new funds this year.

no relation is to be marked between "1995" and "this year".

4.4 Gerunds

Gerunds (verbal forms using a present participle) are not markable. In

Slowing the economy is supported by some Fed officials; *it* is repudiated by others.

one should not mark the relation between "slowing the economy" and "it". A phrase headed by a present participle is taken to be verbal if it can take an object (as in the above example) or can be modified by an adverb.

Present participles which are modified by other nouns or adjectives ("program trading", "excessive spending"), are preceded by an article ("a", "the", "my", etc.) or are followed by an "of" phrase ("slowing of the economy") are to be considered noun-like and ARE markable.

4.5 Pronouns

The possessive forms of pronouns used as determiners are markable. Thus in

its chairperson

there are two potential markables for relations: "its" and the entire NP, "its chairperson". Similarly, in "the man's arm", there are two markables, "[the] man" and "the man's arm". The general question of what is to be treated as a lexical token (apostrophes in this case) is discussed in the MUC document titled "Tokenization Rules."

First, second, and third-person pronouns are all markable, so in

"There is no business reason for *my* departure", *he* added.

"my" and "he" should be marked as coreferential. Reflexive pronouns are markable, so in

He shot *himself* with *his* revolver.

"He", "himself", and "his" should all be marked coreferential. Emphatics are also markable; thus, "himself" should also be marked coreferential, so that "He" and "himself" are marked coreferential in:

He is, *himself*, unsure of the outcome.

In certain cases, pronouns may not have an antecedent ("It's raining") or they may refer to something unmarkable, for example, a clausal construction -- see section 4.1 above.

4.6 Bare Nouns

Prenominal modifiers (e.g., *ocean drilling* in "the ocean drilling company") are markable only if either the prenominal modifier is coreferential with a named entity or to the syntactic head of a maximal noun phrase. That is, there must be one element in the coreference chain that is a head or a name, not a modifier. Thus the following instance is markable, because the prenominal modifier "aluminum" is coreferential with the head noun "aluminum" in the phrase "market for aluminum".

The price of *aluminum* siding has steadily increased, as the market for *aluminum* reacts to the strike in Chile.

Similarly, the following two occurrences of "drug" would be marked:

He was accused of money laundering and *drug* trafficking. However, the trade in *drugs*....

Contrast this with the following occurrences of "contract" and "contract drilling" which would NOT be marked, because there are no occurrences of this phrase, except as a prenominal modifier in the following sequence:

Ocean Drilling & Exploration Co. will sell its *contract drilling* business. ... Ocean Drilling said it will offer 15% to 20% of the *contract drilling* business through an initial public offering in the near future.

Note that the occurrences of *its*, *its contract drilling business* and *the contract drilling business* would all be markable -- see section 6.5.

While nouns in prenominal positions are sometimes markable, the noun which appears at the head of a noun phrase is not separately markable -- it is markable only as part of the entire noun phrase. Thus in the passage

Linguists are a strange bunch. Some linguists even like spinach.

it would not be correct to link the two instances of "linguists".Similarly, in the sentence:

The rate, which was 6 percent, was higher than that offered by the other bank.

the noun phrase "the rate" is a function expression, instantiated by the predicate "6 percent", so these two would be marked coreferential, as follows:

<COREF ID="0" MIN="rate">The rate, which was <COREF ID="1" REF="0">6 percent</COREF>,</COREF> was higher than that offered by the other bank.

In this example, pronoun *that* is coreferential at the FUNCTION level with *The rate*. However, *that* occurs as the head of a noun phrase, *that offered by the other bank*, which is NOT coreferential with *The rate* and *6 percent* (indeed, it refers to a higher rate), so *that* is an instance of a pronoun that cannot be marked in our current framework, even though we lose some type coreference information by not marking it.

4.7 Implicit Pronouns

Assume that English has no zero pronouns; in other words, the empty string is not markable. In

Bill called John and spoke with him for an hour.

there is no relation between the implicit subject of "spoke" and "Bill".

Do not code relations between a relative pronoun and the NP to which it attaches or the gap that it fills. Thus, in

the movie which I saw

the relative pronoun "which" bears no markable relation to either "the movie" (the head to which the relative pronoun attaches) or to the implicit object of "saw" (the gap that the pronoun fills).

4.8 Conjoined Noun Phrases

Noun phrases which contain two or more heads (as defined in section 4.1) are marked by defining the MINimal string (see section 5) as the span from the first "head" through the last "head" including all material in between. The MAXimal string includes the entire maximal conjoined noun phrase. Thus we mark coreference between "The sleepy boys and girls" and "their" as follows:

<COREF ID="1" MIN="boys and girls">The sleepy boys and girls</COREF> enjoy <COREF ID="2" REF="1" TYPE="IDENT">their</COREF> breakfast.

In addition, the individual conjuncts are markable if they are separately coreferential with other phrases:

<COREF ID="1">Edna Fribble</COREF> and <COREF ID="2">Sam Morton</COREF> addressed the meeting yesterday. <COREF ID="3" REF="1" TYPE="IDENT" MIN="Fribble">Ms. Fribble</COREF> discussed coreference, and <COREF ID="4" REF="2" TYPE="IDENT" MIN="Morton">Mr. Morton</COREF> discussed unnamed entities.

<COREF ID="1" MIN="Fribble">Ms. Fribble</COREF> was <COREF ID="2" REF="1" TYPE="IDENT">president</COREF> and <COREF ID="3" REF="1" TYPE="IDENT" MIN="CEO"> CEO of Amalgamated Text Processing Inc.</COREF>

5. HOW MUCH OF THE MARKABLE TO ANNOTATE

The task is defined in order to allow maximal latitude for systems in identifying markables, and to decouple the evaluation from that of accurately parsing noun phrases. Accordingly, the string generated by a system to identify a markable must include the head of the markable (as defined below) and may include any additional text up to a maximal noun phrase (as defined below).

In preparing the key, the text element to be enclosed in SGML tags is the maximal noun phrase; the head will be designated by the MIN attribute.

[We expect that in the future it may be possible, when separate noun phrase bracketings are available, to automatically generate the maximal NP markup from a markup using only heads.]

5.1 Head of a Phrase

For most noun phrases, the head will be the main noun, without its left and right modifiers.

<COREF MIN="task" ...>the coreference task</COREF>

<COREF MIN="contract" ...>the last contract you will ever get</COREF>

<COREF MIN="quantity" ...>a large quantity of sugar</COREF>

<COREF MIN="tons" ...>about 200,000 tons of sugar</COREF>

If the head is a name, the entire name is marked. This includes suffixes such as "Sr.", "III", etc. on personal names and "Corp." on organization names; it does not include personal titles or any modifiers. We follow in this regard the rules for marking personal and organization names for the Named Entity task, as well as for other non-geographic names (e.g, "New Year"):

<COREF MIN="Frederick F. Fernwhistle Jr." ...>the Honorable Frederick F. Fernwhistle Jr.</COREF>

<COREF MIN="Ford Motor Co." ...>Ford Motor Co. of Dearborn, Michigan<COREF>

<COREF MIN="Georg Rath" ...>Herr Dr. Georg Rath</COREF>

In the case of location designators consisting of multiple names, each name is considered a separate unit (as in the Named Entity task) and the head is generally the first of these names, with the others treated as modifiers of the first name:

Beth Sundheim is concerned that this may not be consistent with the MUC Named Entity definition for locations.

<COREF MIN="Newark" ..>Newark, New Jersey</COREF>

Dates, currency amounts, and percentages are also treated as atomic units, as in the Named Entity task:

<COREF MIN="December 7, 1941" ...> December 7, 1941, a day which will live in infamy,</COREF>

<COREF MIN="$1.2 million" ...>$1.2 million in crisp bills</COREF>

<COREF MIN="20%">20% of the shares</COREF>

In the case of "headless" constructions, the "head" -- for coreference purposes -- shall be the last token of the noun phrase preceding any prepositional phrases, relative clauses, and other "right modifiers":

<COREF MIN="seven" ...>seven of the best</COREF>

<COREF MIN="five" ...>the five who were left standing</COREF>

<COREF MIN="youngest"...>the six youngest</COREF>

For constructions that are idioms or collocations, the minimal phrase will ignore the fact that this is a collocation and use the syntactic head; this is because the definition of a collocation is often domain-specific. In the following examples, the MIN is indicated by asterisks:

income *taxes*

light *year*

run of the mill

If the maximal noun phrase is the same as the head, the MIN need not be marked. Also, if the maximal noun phrase differs from the MIN only by the articles "a" or "the", the MIN need not be marked, because the scoring program will automatically strip these before comparing answers.

5.2 Maximal Noun Phrase

The maximal noun phrase includes all text which may be considered a modifier of the noun phrase. This includes (among other modifiers) appositional phrases, non-restrictive relative clauses, and prepositional phrases which may be viewed as modifiers of the noun phrase or of a containing clause:

Mr. Holland

the senior of the executives who will assume Holland's duties

the rumor that the war had ended

Fred Frosty, the ice cream king of Tyson's Corner,

the Penn Central Co., which used to run a railroad,

XYZ Inc. formed *a joint venture with Sony*

Note that in the fourth and fifth cases the final comma may be viewed as part of the NP, and so is included in the maximal NP. The system does not need to worry about punctuation, since the scorer strips punctuation before comparing key to response. In the last case, "with Sony" could equally well be taken to modify "venture" or "formed", and so is included as part of the maximal NP around "venture". Note also that in the "Fred Frosty" example, there is a coreference between the entire noun phrase and the appositional phrase, "the ice cream king of Tyson's Corner"; see section 6.3 for a discussion of this construct. In the case of a conjoined noun phrase with shared complements or modifiers, the maximal noun phrase for the conjoined phrase is the maximal noun phrase. The minimal noun phrase will begin the at the minimal phrase for the first conjunct and include everything up to the end of the minimal phrase for the last conjunct.

If the conjuncts are referenced individually, the maximal noun phrases will NOT include the conjunct. The maximal NP for the first conjunct will include all of the NP up to the conjunction; the maximal NP for the second conjunct will include all of the NP following the conjunction:

<COREF ID="1" MIN="Fribble">Ms. Fribble</COREF> was <COREF ID="2" REF="1" TYPE="IDENT" STATUS="OPT">president</COREF> and <COREF ID="3" REF="1" TYPE="IDENT" STATUS="OPT" MIN="CEO"> CEO of Amalgamated Text Processing Inc.</COREF>

It is possible that the maximal span of the noun phase is interrupted by material that is not part of the noun phrase. Such discontinuous noun phrases should nonetheless be included within a single COREF tag. [In the future, it may be possible to capture the discontinuity explicitly by some special notation.] In the MUC-6 corpus, discontinuous noun phrases frequently appear in headlines, since the non-first lines of a headline are often marked with "@", which is external to the preceding and subsequent text. An annotated example of a discontinuous markable is shown in the example below:

<HL> <COREF ID="0" MIN="Operation">Contract-Drilling Operation

@ in Texas Town</COREF>

@ Sold to Highest Bidder </HL>

In transcripts of spoken languages, a noun phrase may be interrupted by such things as an indication of silence or by the utterance of another speaker. The following two excerpts contain annotated examples of such markables:

<COREF ID="0" MIN="E two">engine <sil> E two</COREF>

<u who=F n=115> I'm going round <COREF ID="0" MIN="mountain">the slate

<u who=G n=116> All the way round

<u who=F n=117> mountain</COREF>

5.3 Exceptions: Articles

If the only difference between the head and the maximal noun phrase is the presence of an article -- the word "the", "a", or "an" at the beginning of the noun phrase -- the MIN need not be explicitly marked. (The scoring program will automatically strip leading articles before comparing strings.)

6. WHICH RELATIONSHIPS TO ANNOTATE

6.1 Basic Coreference

The basic criterion for linking two markables is whether they are coreferential: whether they refer to the same object, set, activity, etc. It is not a requirement that one of the markables is "semantically dependent" on the other, or is an anaphoric phrase.

6.2 Bound Anaphors

We also make a coreference link between a "bound anaphor" and the noun phrase which binds it (even though one may argue that such elements are not coreferential in the usual sense). Thus we would link a quantified noun phrase and a pronoun dependent on that quantification:

Most computational linguists prefer *their* own parsers.

Note that a quantified noun phrase would also be linked to subsequent anaphors, outside the scope of quantification, through the usual relation of identity of coreference. Thus in the following text all three noun phrases would be linked:

Every TV network reported *its* profits yesterday. *They* plan to release full quarterly statements tomorrow.

By this rule, a pronoun in a relative clause which is bound to the head of the clause would get a coreference link to the entire NP. Thus, for

every man who knows his own mind

we would establish a coreference link between "his" and the entire noun phrase "every man who knows his own mind":

<COREF ID="1" MIN="man">every man who knows <COREF ID="2" REF="1" TYPE="IDENT">his <COREF>own mind</COREF>

6.3 Apposition

A typical use of an appositional phrase is to provide an alternative description or name for an object:

Julius Caesar, the well-known emperor,

Julius Caesar, a well-known emperor,

the well known emperor, Julius Caesar,

This identity of reference is to be represented by a coreference link between the appositional phrase, "the well-known emperor" and the ENTIRE noun phrase, "Julius Caesar, the/a well-known emperor":

<COREF ID="1" MIN="Julius Caesar">Julius Caesar, <COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT"> the/a well-known emperor,</COREF></COREF>

The appositional phrase may be separated from the head by other modifiers. Thus

Peter Holland, 45, deputy general manager,...

becomes

<COREF ID="1" MIN="Peter Holland">Peter Holland, 45, <COREF ID="2" REF="1" TYPE="IDENT" MIN="manager"> deputy general manager,</COREF> </COREF>

Appositional phrases are markable (and support the Descriptor slot in the Template Element task in the MUC-7 Information Extraction Task Definition) even when indefinite, e.g.,

Ms. Ima Head, a 10-year MUC veteran,

San Diego, one of America's finest cities,

An appositional phrase is also marked in the specifier relation, e.g.,

<COREF ID="1" MIN="job">The job of <COREF ID="2" REF="1">manager</COREF></ COREF>

However, appositional phrases are NOT marked when they are negative:

Ms. Ima Head, never a great MUC fan,

or when there is only partial overlap of sets:

The criminals, often legal immigrants, ...

Appositional phrases are marked only when they constitute a separate noun phrase following the head. In written text, appositives are generally set off by commas; in transcripts of spoken language, the commas may well not be present because punctuation is generally not captured in text-to-speech transcription.

There are cases where a construction that looks similar to an appositive but occurs within a single noun phrase as a title or modifier, e.g.,

the real estate company * Century 21*

This kind of single noun construction is not considered markable. Thus, no coreference is marked in cases such as the following:

the real estate company Century 21

the realtor Century 21

presidential advisor Joe Smarty

Treasury Secretary Bucks

But the following phrase would have mark-up:

*the job of *manager**

6.4 Predicate Nominals and Time-dependent Identity

Predicate nominals are also typically coreferential with the subject. Thus in the example

Bill Clinton is *the President of the United States*.

we would record a coreference link between "Bill Clinton" and "the President of the United States". Coreference should NOT be recorded if the text only asserts the possibility of identity between two markables. In

Phinneas Flounder may be the dumbest man who ever lived.

Phinneas Flounder was almost the first president of the corporation.

If elected, Phinneas Flounder would be the first Californian in the Oval Office.

no coreference is to be recorded.

We also allow coreference to be recorded when the predicate nominative is marked indefinite, e.g.,:

Mediation is *a viable alternative to bankruptcy*.

Farm-debt mediation is *one of the Farm Belt's success stories*.

ARPA program managers are *nice people*.

However, as with apposition, if there is possibility or a partial set overlap, no coreference is marked because there is no set coreference and no IDENT relation:

Mediation is often a viable alternative to bankruptcy.

Mediation may be a viable alternative to bankruptcy.

Two markables should be recorded as coreferential if the text asserts them to be coreferential at ANY TIME.[2] Thus

Henry Higgins, who was formerly sales director for Sudsy Soaps, became president of Dreamy Detergents

should be annotated as

<COREF ID="1" MIN="Henry Higgins">Henry Higgins, who was formerly <COREF ID="2" MIN="director" REF="1" TYPE="IDENT">sales director for Sudsy Soaps,</COREF></COREF> became <COREF ID="3" MIN="president" REF="1" TYPE="IDENT">president of Dreamy Detergents</COREF>

[2]This is one portion of the guidelines that will clearly need modification after a decision is made about enhancing the notation to distinguish time-dependent coreference links from other coreference links. But the distinction between the two types applies not only to predicate nominals but also to apposition, function-value and other construction types. Thus the phrase "and Time-Dependent Entity" should probably be removed from the title of section 6.4; a new section could cover the general issue of time-dependent coreference. Also, general info about any new type of notation should go in section 2, and the meaning of the new notation should be documented in section 6 (which may need a different section title).

Even if the copula or inchoative verb is embedded, coreference should be marked, as in

Dreamy Detergents named Henry Higgins to be president

which should be annotated as

Dreamy Detergents named <COREF ID="1">Henry Higgins</COREF> to be <COREF ID="2" REF="1" TYPE="IDENT">president</COREF>

When the copula is clearly implied by the semantics of the verb, coreference should be marked. Expressions of equivalence involving the word "as" will also be marked. The NPs enclosed in asterisks in the following examples will be marked coreferential:

Dreamy Detergents named *Henry Higgins* *president*

Henry Higgins is considered *Sudsy Soap's best sales director*

Higgins will serve as *president of Dreamy Detergents*

Cases may arise where an intensional descriptor may apply to two distinct entities, e.g., the current president of a company and a previous president are mentioned. The conventions require that the extensional descriptions guide the mark-up process, and therefore that these two chains *NOT* be collapsed into a single chain:

<COREF ID="1" MIN="Henry Higgins">Henry Higgins, who was formerly <COREF ID="2" MIN="director" REF="1" TYPE="IDENT">sales director for <COREF ID="4">Sudsy Soaps</COREF>,</COREF></COREF> became <COREF ID="3" MIN="president" REF="1" TYPE="IDENT">president of Dreamy Detergents</COREF>. <COREF ID="5" REF="4">Sudsy Soaps</COREF> named <COREF ID="6">Eliza Dolittle</COREF> as <COREF ID="7" MIN="director" REF="6">sales director</COREF> effective last week.

Note that in this example, there will be 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]

Both coreference chains contain the same intensional predicate, "sales director for Sudsy Soaps", but these have different temporal extensional realizations. Although the occurrences of "sales director" are coreferential at the type level, the extensionally grounded chains take precedence, because it is critical to preserve independence of chains grounded in different extensions -- that is, to prevent obviously different individuals from ending up in the same IDENT coreference chain or equivalence class. Thus we "cut" the chain in the above example at the first "type" coreference that would cause collapsing (that is, that can point to a new extension), as in:

<COREF ID="5">Fred</COREF> resigned as <COREF ID="6" MIN="president" REF="5">president of IBM</COREF>; next month, <COREF ID="7">the president</COREF> will be <COREF ID="8" REF="7">Mary</COREF>.

6.5 Types and Tokens

The general principle for annotating coreference is that two markables are coreferential if they both refer to sets, and the sets are identical, or they both refer to types, and the types are identical. There are a number of problematic cases where one can argue whether something is a set or a type. There is no simple algorithm for determining the ontological category of a referent. There are, though, some useful rules. Most occurrences of bare plurals refer to types or kinds, not to sets. In

...*producers* don't like to see a hit wine increase in price... *Producers* have seen this market opening up and *they*'re now creating wines that appeal to these people.

"producers", "Producers", and "they" refer to types and they all refer to the same type. Notice that if interpreted as referring to sets, they would not all refer to the same set. More properly, there is no reason to think they would corefer; not all the producers who have seen the market opening up have created new wines.

Note that a type can be referred to by a bare plural, a definite singular NP ("the tiger is fast becoming extinct") or a (bare) prenominal. In

The action followed by one day an Intelogic announcement that it will retain an investment banker to explore alternatives "to maximize *shareholder* value," including the possible sale of the company. Mr. Edelman declined to specify what prompted the recent moves, saying they are meant only to benefit *shareholders* when "the company is on a roll."

the two starred occurrences corefer to the type: shareholder (of Intelogic).

6.6 Functions and Values

In

GM announced *its third quarter profit*. *It* was *$0.02*.

all three starred phrases refer to an amount of money; they all refer to the same amount of money. Hence they are coreferential. The first phrase, in context, refers to that amount via referring to a function, say of companies and quarters of a year--or times. (In addition, the "its" in the first NP would be linked to GM.) In

General Motors announced {their third quarter profit of *$0.02*}.

the bracketed and starred phrases are coreferential. They refer to one and the same amount of money. Note that here, as in the case of apposition, the result is that a phrase is marked as being coreferential with a part of the phrase.

In

The temperature is *90*....The temperature is rising.

the first occurrence of "the temperature" is an intensional expression referring to the value (extension) of the function at arguments (places, times) supplied by context. *The temperature* is coreferential with "90" which grounds it. In the second occurrence, "the temperature" refers to the function (indirectly, by way of referring to the derivative of the function). So it is not coreferential with the first occurrence or with "90".

In the sequence

The temperature was 90 yesterday and has already reached 95 today. This sets a new record high.

we have a different problem: we have two extensional descriptions (90, 95) for the temperature, and only a single occurrence of the intensional description "temperature". In this case, "temperature" is coreferential with the extensional description occurring in the same clause ("90"). As a result, "95" is in its own coreference class, and we are not able to mark the fact that it too is a temperature. However, "95" is coreferential with "This" and "a new record high". This is marked as follows:

<COREF ID="4">The temperature</COREF> was <COREF ID="5" REF="4">90</ COREF> yesterday and has already reached <COREF ID="6">95</COREF> today. <COREF ID="7" REF="6">This</COREF> sets <COREF ID="8" MIN="high" REF="7">a new record high</COREF>.

If both extensions are in the same clause, as in:

The stock value rose from $8.05 to $9.15

The per share value of $8.05 rose to $9.15 at the end of trading.

then the function takes on the most "current" value in its clause, e.g., *stock value* and *$9.15* are marked coreferential, and $8.05 is in its own class, not coreferential.

6.7 Metonymy

The pervasive phenomenon of metonymy raises a problem for Coreference relations. Do we annotate and

recognize the relation before or after coercion? Here are some texts to consider:

(1) *The White House* sent its health care proposal to Congress yesterday. Senator Dole said *the administration*'s bill had little chance of passing.

(2) *Ford* announced a new product line yesterday. *Ford* spokesman John Smith said *they* will start manufacturing widgets.

(3) I bought the New York Times this morning. I read that the editor of the New York Times is resigning.

(4) *The United States* is a democracy. *The United States* has an area of 3.5 million square miles.

We propose that coreference be determined with respect to coerced entities. Of course, this still leaves open the question as to the circumstances under which coercion is required. In (1) there is a coercion from the White House to the administration operating out of the White House, and that is IDENT with "the administration"; so "White House" and "administration" are IDENT. (Notice that there is also a question as to whether the adminstration's proposal is the same as its bill. This too requires a coercion of sorts.) In (2), while there might seem to be a coercion from Ford to a spokesman for Ford, we believe that such a coercion is not necessary, for it is plausible that corporations, as legal persons, can do many of the things that people can do--such as "announce." They may have to do some or all such things through other agents, but many people do many things that way. And if Ford can announce, then it, through one of its spokesmen, can "say". Believing that no coercion is required, we would mark as coreferential the first instance of "Ford", the second instance of "Ford" (in the phrase "Ford spokesman John Smith"), and "they", but would NOT mark the phrase "Ford spokesman John Smith" as coreferential with anything else in this passage. In (3) the first "New York Times" is coerced into a copy of the paper published by the New York Times and the second is coerced into the organization; so they are not IDENT. (4) is somewhat akin to (2). Countries are both geographical entities and governmental units. Thus, no coercion is necessary and the two starred occurrences are coreferential.

In the absence of general principles, a body of such decisions will need to be developed to codify the rules for coercion and coreference. In cases where there has been no clear precedent, the answer keys for formal evaluations will need to mark coreference as optional.

7. BASIS OF JUDGMENT

The coreference judgments should be based on the intelligent reader's knowledge of the world resulting from his or her best understanding of the text. It should not be based on a theory of the structure of the text, or on a linguistic theory of how NPs are resolved, or on estimates of what the typical NLP system could do. This means that some relations will be impossible for current NLP systems to recover, but this is why the task will push the technology. The annotators should assume that they are typical intelligent readers.

8. SCORING AND THE ORDERING OF LINKS

If three markables, A, B, and C, are coreferential, this relationship could be recorded in the key in several ways: for example, by a REF pointer in both B and C pointing to A, or by a REF pointer in B pointing to A and a REF pointer in C pointing to B. A similar range of variations is possible in a system response. The current scoring rules provide that any correct key, when compared to any correct response, will yield a 100% recall/100% precision score, independent of the way the coreference relation is encoded in the key by REF pointers.

APPENDIX A. SAMPLE ANNOTATIONS

This appendix contains annotations as they would appear in an answer key. The notation produced by a coreference identification system would differ from the answer-key notation in certain respects (see information in the guidelines concerning the MIN and STATUS attributes).

A.1 Sentences Excerpted from a MUC-6 Wall Street Journal Article

<s> <COREF ID="0">Ocean Drilling & Exploration Co.</COREF> will sell <COREF ID="3" MIN="business"><COREF ID="2" TYPE="IDENT" REF="0">its</ COREF>contract-drilling business</COREF>, and took a $50.9 million loss from discontinued operations in <COREF ID="12" MIN="quarter">the third

quarter</COREF> because of the planned sale. </s>

<s> <COREF ID="9" TYPE="IDENT" REF="2" MIN="company">The New Orleans oil and gas exploration and diving operations company</COREF> added that <COREF ID="10" TYPE="IDENT" REF="9">it</COREF> doesn't expect any further adverse financial impact from the restructuring. </s>

…

<s> In <COREF ID="11" TYPE="IDENT" REF="12" MIN="quarter">the third quarter</COREF>, <COREF ID="13" TYPE="IDENT" REF="10" MIN="company">the company, which is 61%-owned by Murphy Oil Corp. of Arkansas,</COREF> had <COREF ID="100" MIN="loss">a net loss of <COREF ID="17" TYPE="IDENT" REF="100">$46.9 million</COREF>, or <COREF ID="16" TYPE="IDENT" REF="17" MIN="91 cents">91 cents a share</COREF>. </s>

…

<s> It has long been rumored that <COREF ID="28" TYPE="IDENT" REF="13">Ocean Drilling</COREF> would sell <COREF ID="29" TYPE="IDENT" REF="3">the unit</COREF> to concentrate on <COREF ID="30" TYPE="IDENT" REF="28">its</COREF> core oil and gas business. </s>

A.2 Transcript of a TRAINS Dialogue

utt1 : s: hello can I help <COREF ID="1">you</COREF>

utt2 : u: yes <COREF ID="0" TYPE="IDENT" REF="1">I</COREF>'d like <sil> to <sil> take <COREF ID="3">a tanker</COREF> from <COREF ID="12">Corning</COREF> and bring <COREF ID="2" TYPE="IDENT" REF="3">it</COREF> to <COREF ID="5">Elmira</COREF>

utt3 : s: alright

utt4 : u: and from <COREF ID="4" TYPE="IDENT" REF="5">Elmira</COREF> <COREF ID="6" TYPE="IDENT" REF="0">I</COREF>'d like to load <COREF ID="17" MIN="juice">orange juice</COREF> <sil> into <COREF ID="7" TYPE="IDENT" REF="2">the tanker</COREF>

utt5 : s: mm-hm

utt6 : u: <COREF ID="8" TYPE="IDENT" REF="6">I</COREF>'d like then to take the el- <sil> <COREF ID="9" TYPE="IDENT" REF="7">the tanker</COREF> back from <COREF ID="10" TYPE="IDENT" REF="4">Elmira</ COREF> to <COREF ID="11" TYPE="IDENT" REF="12">Corning</COREF>

utt7 : s: alright

utt8 : u: now from <COREF ID="13" TYPE="IDENT" REF="11">Corning</COREF> what would be the quickest route to <COREF ID="23">Avon</COREF>

utt9 : s: uh through <COREF ID="21">Dansville</COREF>

utt10 : u: okay then <COREF ID="14" TYPE="IDENT" REF="8">I</COREF>'d like to take <sil> <COREF ID="15" TYPE="IDENT" REF="9" MIN="tanker">the <sil> tanker of <COREF ID="16" TYPE="IDENT" REF="17" MIN="juice">orange juice</ COREF></COREF> through <COREF ID="20" TYPE="IDENT" REF="21">Dansville</ COREF> and then on to <COREF ID="22" TYPE="IDENT" REF="23">Avon</COREF>

utt11 : s: alright <sil> um which engine would <COREF ID="24" TYPE="IDENT" REF="14">you</COREF> like to use

utt12 : u: <COREF ID="26" MIN="E two">engine <sil> E two</COREF>

utt13 : s: alright

utt14 : u: well is is <COREF ID="25" TYPE="IDENT" REF="26">E two</COREF> which is a <sil> will either one take <COREF ID="27" TYPE="IDENT" REF="24">me</COREF> there any quicker

utt15 : s: uh no <COREF ID="101">they</COREF>'re a <COREF ID="100" TYPE="IDENT" REF="101">they</COREF>'re both the same

utt16 : u: okay then <COREF ID="28" TYPE="IDENT" REF="27">I</COREF>'ll take <COREF ID="29" TYPE="IDENT" REF="25" MIN="E two">engine E two</ COREF>

utt17 : s: alright

utt18 : u: and the tankers move at the same speed

utt19 : s: right

utt20 : u: okay

utt21 : s: so uh the plan is to take <brth> <sil> <COREF ID="30" TYPE="IDENT" REF="29" MIN="E two">engine E two</COREF> <sil> to <COREF ID="31" TYPE="IDENT" REF="13">Corning</COREF> pick up <COREF ID="32" TYPE="IDENT" REF="7">a tanker</COREF> <sil> back to <COREF ID="33" TYPE="IDENT" REF="10">Elmira</COREF>

utt22 : um <sil> load <COREF ID="34" TYPE="IDENT" REF="32">it</COREF> with <COREF ID="35" TYPE="IDENT" REF="16" MIN="juice">orange juice</ COREF> <sil> and then <sil> to <sil> <COREF ID="37" TYPE="IDENT" REF="22">Avon</COREF>

utt23 : u: by way of <sil> + <COREF ID="38" TYPE="IDENT" REF="31">Corning</COREF> + <sil> and <COREF ID="39" TYPE="IDENT" REF="20">Dansville</COREF>

utt24 : s: + by way of +

utt25 : right

utt26 : u: how long will that take

utt27 : s: eleven hours

utt28 : u: okay <sil> <COREF ID="40" TYPE="IDENT" REF="28">I</COREF> am now finished

---

Last Modified August 2000