File name: co-task-def

Coreference Task Definition

(V2.3, 8 Sep 95)

---

1 GENERAL NOTATION

1.1 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.

1.2 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.

1.3 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.

1.4 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 4 for a full discussion of this issue.

1.5 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"). 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.

```
<COREF ID="102" MIN="Board of Education">Our 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.
```

2 WHAT PART OF THE TEXT TO ANNOTATE

The <TXT> portion of the article should be annotated as well as the <HL>, the <DD>, and the <DATELINE> from the article header, but not any other lines from the header. (The DD tag sometimes doesn't appear at all, sometimes appears once, and sometimes appears twice. When it appears twice, only the SECOND instance is to be annotated.)

Lines within the <TXT> portion of the article that start with the "@" sign signify a table or other special line formatting within the text and should NOT be annotated. (However, such lines may also appear within the <HL> portion of the article, and these should be annotated.)

3 WHAT THINGS TO ANNOTATE

3.1 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.

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".

Some indefinite NPs are not markables. See section 5.

3.2 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. 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".

3.3 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 "the" or are followed by an "of" phrase ("the slowing of the economy") are to be considered noun-like and ARE markable.

3.4 Pronouns

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

> its chairperson

there are are two potential markables for relations: "its" and the entire NP, "its chairperson". Similarly, in "the man's arm", there are two markables.

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.

3.5 Bare Nouns

Prenominal occurrences of nouns, e.g., in compound nouns, are markable. Thus in

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

the relation between the two occurrences of "aluminum" should be marked. Note this presupposes that the two occurrences co-refer; they do, they both refer to the type of material.

While nouns in prenominal positions are 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".

3.6 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 head it attaches to or the gap that it fills.

3.7 Conjoined Noun Phrases

Noun phrases which contain two or more heads (as defined in section 4.1) are NOT markable. This restriction is imposed so that each markable can be identified by a unique contiguous head substring. Thus no coreference is to be marked for

> The boys and girls enjoy their breakfast.

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.
```

If the conjuncts share modifiers, the coreference is optional:

```
<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>
```

4 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.]

4.1 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</COREF> you will ever get
<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.

```
<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:

```
<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>
```

If the maximal noun phrase is the same as the head, the MIN need not be marked.

4.2 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; 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 5.3 for a discussion of this construct.

In the case of a pair of conjoined noun phrases with shared complements or modifiers, 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>
```

4.3 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.)

5 WHICH RELATIONSHIPS TO ANNOTATE

5.1 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.

5.2 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>
```

5.3 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,

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 well-known emperor":

```
<COREF ID="1" MIN="Julius Caesar">Julius Caesar, <COREF ID="2" REF="1"
   MIN="emperor" TYPE="IDENT"> the 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 that are marked indefinite are NOT considered to be coreferential. Examples of noncoreferential appositional phrases include the following:

Ms. Ima Head, a 10-year MUC veteran,
San Diego, one of America's finest cities,

Currently, only appositional phrases that are overtly marked via punctuation are considered markables. 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*
*the job of *manager**

5.4 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 is a leading candidate to become president.
If elected, Phinneas Flounder would be the first Californian in the Oval Office.

no coreference is to be recorded.

Neither should coreference be recorded when the predicate nominative is marked indefinite. Examples of noncoreferential predicate nominatives include

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.

Two markables should be recorded as coreferential if the text asserts them to be coreferential at ANY TIME. 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>
```

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 implied by the semantics of the verb but is not expressed overtly, the coreference relation will be marked optional in the answer key. Expressions of equivalence involving the word "as" will also be marked optional. The NPs enclosed in asterisks in the following examples will be marked optionally 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*

5.5 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).

5.6 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" refers to the value of the function at arguments (places, times) supplied by context. That occurrence is coreferential with "90". 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".

There will be cases where a phrase could arguably refer to either a set or a type; in such ambiguous cases, the coreference should be recorded but marked as optional.

5.7 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.

6 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.

7 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. However, if the response is incomplete, its recall score CAN be affected by the way in which the coreference relation is encoded by the key. It is therefore recommended that each markable which participates in a coreference relation have a REF pointer to the most recent prior coreferential markable which does not have STATUS="OPT".