The MATE meta-scheme for coreference in dialogues in multiple languages

M. Poesio[+], *F. Bruneseaux*,[†] *and L. Romary*[†]

[+]HCRC and CSTR, University of Edinburgh, Edinburgh, Scotland
{Massimo.Poesio}@ed.ac.uk
[†]LORIA, Nancy, France
{brunesea,romary}@loria.fr

1 Introduction

One of the goals of the EU-funded MATE project is to develop tools to support some of the most popular dialogue coding activities, including annotation of syntactic information, information about 'coreference,' and information about dialogue acts.[1]

A problem to be confronted when trying to develop such tools is the lack of universally agreed upon coding schemes for these 'levels'[2]–i.e., of specifications of a set of elements and attributes that will cover all of the information about that level that a researcher may wish to annotate, together with instructions for how to do so. What does exist at the moment is coding schemes for particular domains and/or applications: in the case of dialogue acts, for example, there are several 'specific' schemes for given applications, some of which have been shown to lead to reliable coding (Alexandersson et al., 1997; Carletta et al., 1997). Recently, attempts to come up with standards for a few of these levels have been made at workshops organized by the Discourse Resource Initiative (DRI). The DRI did come up with a proposal concerning the dialogue act level (Discourse Resource Initiative, 1997; Allen and Core, 1997), although there have been serious disagreements concerning the usefulness of such a 'standard' for this level, since it's not clear that it's possible to come up with a domain-independent definition of dialogue acts. No official recommendation has been made in the DRI for the so-called 'coreference' level, although the DRAMA scheme (Passonneau, 1997) has sometimes been discussed for this purpose.

In this paper we report on the current proposal concerning the type of 'coreference' annotation to be supported by the MATE workbench, motivating our proposal by relation to previous proposals in this area. The full proposal is available on-line at http://www.cogsci.ed.ac.uk/~poesio/MATE/coreference_scheme.html.

[1]The project's home page is at http://mate.nis.sdu.dk/.

[2]We use the term MARKUP LEVEL to refer to each of these types of annotation.

2 Annotating for 'Coreference'

2.1 Problems to be addressed

The difficulties to be addressed in the case of the 'coreference' level are of a different nature from those that arise in the case of the other annotation task at the semantic level considered in MATE, the dialogue acts level.

A very basic problem arising in the case of coreference is deciding what type of information is being annotated, since the term 'coreference' is used to indicate different things. The name 'coreference' derives from the task of 'coreference resolution', one of the semantic interpretation tasks adopted in the Message Understanding Conference (MUC), a US initiative to evaluate systems performing information extraction. The coreference annotation scheme used in MUC-7, MUCCS (Hirschman, 1997) was devised to evaluate the ability of the systems participating in the competition to identify which elements in the text referred to the same object; hence the term 'coreference'. The scheme adopted for annotating references to landmarks in the MapTask corpus is also meant to annotate reference in this sense. Both the DRAMA scheme and the schemes proposed by Lancaster University (Fligelstone, 1992), instead, are meant to be used to annotate anaphoric information in texts; but coreference is not the same as anaphoricity. Two NPs can corefer without either of them being 'anaphoric' in the traditional sense –e.g., proper names are not generally considered 'anaphoric expressions', yet two proper names can obviously corefer, as in (1a); and conversely, two NPs can be in an anaphoric relation without either of them 'referring' to anything, as in (1b), where *one of the engines at Elmira* doesn't really refer to any specific engine yet serves as antecedent of *that*. (See

(van Deemter and Kibble, 1999) for further discussion.) Fortunately for our purposes, coreference information can be expressed using the relations used to express anaphoric information, which makes it possible to develop schemes in which both types of information can be encoded, as we will see below. We will keep referring to the markup level with which we are concerned as 'coreference' for consistency with common use, but the reader should use the term with care.

(1) a. ... home fans at the Stade de France endured an agonising final 20 minutes after Laurent Blanc was shown the red card following a tussle with Slaven Bilic. Blanc ...

b. 15.1 M: +we're+ gonna hook up ONE OF THE ENGINES AT ELMIRA to the boxcar at Elmira
15.6 : shove THAT off to Corning

One advantage of the coreference level is that notwithstanding the possible source of confusion just mentioned, there is much more agreement on the underlying catalogue of semantic notions needed to characterize this type of information than there is, say, for the discourse act level, so that it's possible to come up with fairly precise definitions of most of the information one would want to annotate. So, the main problem the designer of such a scheme has to confront is the sheer pervasiveness of the phenomenon: almost every word in a coherent text–including quantifiers, nouns, (modal) verbs, and adjectives–can be said to be anaphorically related in some way to what has already been introduced in the text, as shown by the following examples. This means that, in practice, it will always be necessary to restrict somehow the amount of anaphoric information to annotate. (See also the discussion in (Hirschman, 1997).)

(2) a. A group of students entered a pub. Three boys ordered beer, ...

b. ...It is in such places that we find some of the most beautiful and adventurous modern architecture, and some of the most intriguing attempts to deepen the experience of art. Newhouse has probably seen more of the recently built art museums ...than anyone else.

A second problem is that we do not know yet what can be annotated reliably and what cannot. In their reliability study, (Poesio and Vieira, 1998) found a fair agreement among annotators (K = .76) concerning which NPs were anaphoric and which ones were not, and about 95% agreement on antecedents for those definite descriptions that all subjects identified as anaphoric; but no agreement on identifying bridging references (K=.24), and often different antecedents were indicated for those that all annotators classified as bridges.[3] As a result, the only coding scheme whose reliability has been extensively tested is MUCCS.

2.2 Existing schemes

The MUCCS scheme (Hirschman, 1997) is the most widely used of the existing coreference schemes, and also the more modest in scope: it concentrates on identity relations between NPs. The main problem with MUCCS from our point of view is that it was designed for texts, so it does not provide instructions either for dealing with typical problems in dialogue such as disfluencies, or for annotating references to the visual situation, common e.g., in the MapTask corpus and in multimodal applications, and that we hypothesize can be reliably annotated (although this hypothesis will have to be verified). Also, it's only designed for English, and therefore does not include instructions for anaphoric expressions common in other European languages and whose relation with other discourse entities could be annotated reliably, such as clitics.

The DRAMA scheme (Passonneau, 1997) does include instructions for dealing with some difficult problems of markable identification in dialogues, but not for multilingual annotation. DRAMA includes instructions for annotating bridging references (whose reliability, however, still has to be ascertained), but not for references to the visual situation. The scheme proposed by (Bruneseaux and Romary, 1998) provides markup elements (based on the TEI scheme) to annotate both references to the visual situation and discourse deixis, in addition to bridging references; the reliability of this type of annotation wasn't evaluated. The Lancaster scheme (Fligelstone, 1992) was also designed for texts, and in certain ways is more ambitious than any of the schemes discussed

[3]Informal studies conducted in MUC and by the DRI confirm this (Discourse Resource Initiative, 1997). The participants to these initiatives found reasonable agreement on the antecedents of anaphoric expressions, but poor recall for bridges. These results led to the elimination of relations other than IDENT from the MUC coding scheme for coreference, MUCCS. Those studies also suggest that one way to improve reliability on bridges may be to use better tools suggesting relations. Also, using a smaller number of relations may help, but the actual number of anaphoric elements annotated was not indicated so it's not clear whether this result is actually significant.

here in that it also contains instructions for annotating elliptical references. We are not aware of any results about the reliability of the scheme.

3 The MATE Proposal

3.1 Approach: A Meta Scheme and Two Instantiations

It should be clear from the considerations above that we do not believe that there is such a thing as a universally useful standard for 'coreference' annotation in dialogues. At the same time, because the semantics of anaphora and coreference is relatively well-understood, it is possible to extract from the schemes discussed above a fairly short list of options available to the designer of a scheme. (This is unlike the case of dialogue acts, where different schemes are very difficult to compare.) These considerations suggested a 'meta-scheme' approach to the goal of developing a scheme for the coreference level that could be useful for different types of applications. What this means is that instead of proposing a single scheme, we identified a range of possible types of information about 'coreference' one may want to annotate, on the basis of the coding schemes for coreference discussed above; we evaluated how reliable each type of annotation is likely to be; and we specified the markup language needed to pursue each option.[4]

The meta scheme consists of a CORE SCHEME and three extensions. The core scheme can be used to do the type of annotation that can be done with MUCCS (i.e., identity relations between discourse entities introduced by NPs). The three extensions to the core scheme can be used to annotate (i) references to the visual situation, as in the Bruneseaux and Romary scheme and in the MapTask scheme for annotating references to landmarks; (ii) a more complex set of relations between entities ('bridges'), as in the DRAMA scheme; and (iii) anaphoric relations involving an extended range of anaphoric expressions (such as incorporated clitics) and of antecedents (as in discourse deixis). The tool will support the whole range of elements and attributes of the meta scheme; the task of the designer of a scheme for a particular application will be to identify the options of interest among those supported by the tool, ignoring the rest. The documentation for the coreference level includes, in additions to a discussion of the core scheme and the extensions, an example of how to extract a scheme from the meta scheme (the example covers references to the visual situation) as well as instructions for using the markup definitions provided by the meta scheme to encode according to the DRAMA scheme.

On the assumption that the designer of a scheme for dialogues may be interested in annotating both 'anaphoric' and 'coreferential' information, we addressed the problem of the difference between the two types of annotation by adopting a position analogous to that taken in DRT (Kamp, 1981; Heim, 1982), whereby coreference information is expressed in terms of the same semantic relations used to annotate anaphoric information. This is done by introducing in the annotation identifiers that stand for the 'actual' referents, and expressing reference by means of relations between the discourse entities that 'refer' and these identifiers.

3.2 Markup Language and Assumptions About File Organization

In the rest of the paper we discuss first the Core Scheme, then each of the extensions. In this section we introduce the markup language and discuss a few assumptions underlying annotation using the MATE Workbench common to several levels, as well as a few assumptions specific to the coreference level.

The markup language in MATE is XML, a simplification of SGML meant to make for easier parsing, but the workbench will make this transparent to annotators (not to annotation scheme designers, of course). An assumption common to most levels is the distinction between BASE FILE and ANNOTATION FILE. The base file contains the information necessary to annotate; the results of the annotation for a given level go in a separate file containing pointers to elements of the base file. For the coreference level, for example, the base file could be either a file annotated with one XML element per word, as in the British National Corpus, or possibly a file containing syntactic information; in the latter case, the text elements to be annotated could be identified automatically.

We assume here that the base file is encoded according to the recommendations for the morpho-syntactic level of chunks adopted in MATE (Pirrelli and Soria, 1999), which specify a type of syntactic representation that could be produced by existing parsers; such parsers might be integrated in the workbench. For example, the representation in terms of chunks of the sentence *John likes Bill* would be as follows:

(3)
```
ch.xml
<ch id="ch_001" type="N">
<potgov id="p_001">John</potgov>
</ch>
```

[4] The 'meta-scheme' approach was also adopted in EAGLES `http://www.ilc.pi.cnr.it/EAGLES/home.html` and the CES `http://www.cs.vassar.edu/CES/`.

```
<ch id="ch_002" type="V">
<potgov id="p_002">likes </potgov>
</ch>
<ch id="ch_003" type="N">
<potgov id="p_003">Bill </potgov>
</ch>
```

The result of an annotation at the coreference level of the file `ch.xml` (the base file) would be to produce a second file containing elements that point to `ch.xml`, as discussed below.

4 The Core Scheme

The Core Scheme has been designed to annotate the subset of all information about 'coreference' that can be annotated reliably: that is, information about identity relations between discourse entities explicitly introduced by elements of a text. As in MUCCS, it is assumed that annotation will proceed in two steps: first agreeing on the markables, then markup of anaphoric relations. The main difference from the MUC scheme is that following the recommendations of the Text Encoding Initiative and of Bruneseaux and Romary, the distinction between these two steps of annotation is mirrored in the Core Scheme by a distinction between two elements: `<de>`, used to mark the parts of text that may be involved in these relations, and `<link>`, used to mark information about these relations.

4.1 Identifying and Marking Discourse Entities

As mentioned above, the main problem with designing a scheme for anaphora and coreference is not that the relevant notions are difficult to define, but that you can't annotate everything. Of the schemes we examined, the Lancaster University is the most ambitious, including also ways for annotating VP ellipsis. DRAMA recommends to annotate all noun phrases, whether or not they introduce discourse entities. MUCCS recommends to annotate also noun phrases occurring in prenominal position in other noun phrases, such as *Getty* in *the Getty museum*.

For the Core Scheme we adopted a conservative view close to that of DRAMA, and only recommend to mark the potential antecedents of anaphoric and referential expressions that are realized in the text as full NPs; in other words, we did not include instructions for annotating parts of NPs that may enter in such relations (as in MUCCS) and for annotating verbal constituents that enter e.g., in ellipsis (as in the Lancaster scheme). If needed, elements for marking up verbal constituents are included in one of the extensions of the Core Scheme (see below), so the MATE Workbench could be used to support this type of annotation, as wellprovided that the users come up with their own instructions for identifying the markables.

Each markable NP is annotated with a `<de>` element with an `ID` attribute. In the underlying XML representation, the `<de>` elements include pointers to elements in the base file; e.g., the markables in the example in (3) would be represented as follows:

(4)
```
coref.xml

<de id="de_001" href="ch.xml#id(ch_001)"/>
<de id="de_002" href=" ch.xml#id(ch_003)"/>
```

However, this aspect of the representation should be transparent to the annotator, who will only be concerned with marking `<de>` elements and assigning them an `ID`. So in what follows, also to make the notation more readable, we will represent markup using a simpler notation without `HREF` pointers, as in the following examples:

(5)
```
<de ID="de_01">we</de>'re gonna take
<de ID="de_07">the engine E3</de>
and shove <de ID="de_08"> it </de> over
to <de ID="de_02">Corning</de>,
hook <de ID="de_09"> it </de> up to
<de ID="de_03">the tanker car</de>...
```

(6)
```
197 F: mmh / Donc qu'est ce que vous
       allez garder en fait (?) + /
198 M: |<de ID="de_96">la longueur du
       <de ID="de_97">tube</de></de>
    et <de ID="de_98"> les ailerons </de>
199 D:<de ID="de_99"> les ailerons </de>
200 F: Donc <de ID="de_100"> les ailerons
       </de> vous m'avez dit.
```

It is assumed that in most cases (at least, when the base file is annotated with syntactic information) markables will be automatically identified by means of search patterns formulated in terms of the MATE query language (Heid and Mengel, 1999); the main role of the annotator would be to correct possible problems. This suggests that the markables would be mostly identified on purely structural grounds. The instructions for identifying markables do include however a discussion of several cases in which the designers of a scheme may decide not to mark a text element as `<de>` even if syntactically it counts as a NP: examples are NPs in predicative position, such as *a policeman* in *John is a policeman*[5], and repeated NPs in the case of disfluencies, as in the following example:

(7)
```
193 F: Donc qu'est ce qui /
       qu'est ce qui serait commun a
```

[5]Note that the pronoun *he* in the continuation *He works for the 27th district* would not be considered ambiguous.

```
        <de>ces deux fusees</de>.
        <de>Ces deux fusees</de> ont /
194 D: c'est qu'elles ont /
        <de>elles</de> ont la meme /
        elles / elles / toutes les /
        tous les ailerons
```

(in cases like this, DRAMA recommends to mark up all repetitions of *elle*, although again, they do not create ambiguity). The instructions for the Core Scheme include a fairly extensive discussion of which text constituents count as NPs, which incorporates examples from MUCCS and DRAMA as well as from (Quirk and Greenbaum, 1973).

An issue not considered either in MUCCS or in DRAMA is what to do when a discourse entity is not introduced by a single contiguous prase, but by utterances interrupted by disfluencies or comments, as in (8), where the utterance of *the diamond mine* is interrupted by an acknowledgment from the follower:

(8)
```
GIVER: curving, just curving round
       the diamond
FOLLOWER: uh-huh
GIVER: mine..... uh-huh
```

We believe this problem should be addressed at the parsing level by providing ways of representing non-contiguous syntactic elements, as done in the representation for the morpho-syntactic level proposed in MATE. The chunk-level representation of the example above is shown in (9), whereas the representation at the coref level is shown in (10).

(9)
```
ch.xml:

GIVER: Curving, just curving round
<ch ID="ch_66" next="ch_68">the diamond</ch >
FOLLOWER: <ch ID="ch_67">uh-huh </ch>
GIVER: <ch ID="ch_68" prev="ch_66">
       mine </ch >.
<ch ID="ch_69">..... uh-huh</ch>
```

(10)
```
coref.xml:
<de ID="DE_01"
    href="ch.xml#id(ch_66)..id(ch_68)"/>
```

The other addition to the instructions given in MUCCS and DRAMA are instructions for marking up clitics and empty elements, common in Italian and Spanish. Markup elements for marking incorporated clitics (such as *daselo* in (11) and empty elements are discussed below; clitics realized as distinct particles (such as *la* in (11) are also marked as <de>s, as follows.

(11)
```
Mira, te doy <de ID="de_167"> este libro </de>
Conoces a <de ID="de_168"> mi suegra?</de>
Pues daselo cuando
<de ID="de_170"> la </de> veas.
```

4.2 Links

The subset of 'coreference' information which has been shown most clearly to be markable in a reliable way coincides with the information that can be annotated with the MUCCS scheme: identity relations between discourse entities explicitly introduced in the text by nominal phrases. In the core scheme, this is the only information that can be annotated.

Whereas identity relations are represented in MUCCS and DRAMA by means of attributes on elements that correspond to the <de> element used in the MATE scheme, we adopted a notation derived from the <link> mechanism used in the TEI for linking any text element and adopted by Bruneseaux and Romary for representing anaphoric information.[6] <link> elements have two attributes: a HREF pointer to the <de> element that stands in an anaphoric relation with an antecedent, and a TYPE attribute specifying the relation (which in the case of the Core Scheme can only be IDENT). <link> elements contain then one or more <anchor> elements, with a single <href> pointer to the antecedent. So for example, the anaphoric relations in (5) and (6) would be annotated as follows:

(12)
```
coref.xml

<de ID="de_01">we</de>'re gonna take
<de ID="de_07"> the engine E3 </de>
and shove <de ID="de_08"> it </de> over
to <de ID="de_02">Corning</de>,
hook <de ID="de_09"> it </de> up to
<de ID="de_03">the tanker car</de>...
<link href="coref.xml#id(de_07)"
      type="ident">
  <anchor href="coref.xml#id(de_08)"/>
</link>
<link href="coref.xml#id(de_08)"
      type="ident">
  <anchor href="coref.xml#id(de_09)"/>
</link>
```

(13)
```
coref.xml:

197 F: mmh / Donc qu'est ce que vous
       allez garder en fait (?) + /
198 M: |<de ID="de_96">la longueur du
       <de ID="de_97">tube</de></de>
       et
       <de ID="de_98">les ailerons</de>
199 D:<de ID="de_99"> les ailerons </de>
200 F: Donc
       <de ID="de_100">les ailerons</de>
       vous m'avez dit.
<link href="coref.xml#id(de_98)"
      type="ident">
  <anchor href="coref.xml#id(de_99)"/>
</link>
<link href="coref.xml#id(de_99)"
      type="ident">
  <anchor href="coref.xml#id(de_100)"/>
```

[6] The slightly modified representation of <link>s in our proposal was adopted for technical reasons.

```
</link>
```

In MUCCS, the annotator is free to choose any of the two elements in an identity relation as 'anaphor', because identity is symmetric. As the intention is to use <link> elements to also annotate non-symmetric relations such as those found in bridging cases, we recommend to always have the HREF pointer in the <link> element point to the anaphor, and the HREF pointer in the <anchor> to the antecedent. (In general, the annotator will still be able to exploit the transitivity property of identity and choose any antecedent, although in particular cases this may not be a good idea either.)

The reason why <link> elements may have more than one <anchor> element is to annotate ambiguities, which are very common in spoken dialogue. In case more than one <de> element appears to be an equally likely antecedent of an anaphoric expression, each of the possibilities should be marked by means of a separate <anchor> element. In (14a), for example, the pronun *it* in 15.16 could refer equally well to engine E3 or the tanker car. Both antecedents should be annotated, as shown in (14b).

(14) a.
```
15.12  : we're gonna take the engine E3
15.13  : and shove it over to Corning
15.14  : hook it up to the tanker car
15.15  : _and_
15.16  : and send it back to Elmira
```

b.
```
coref.xml:

15.12  : we're gonna take
           <de ID="de_15">the engine E3</de>
15.13  : and shove <de ID="de_16"> it </de>
           over to Corning
15.14  : hook <de ID="de_17">it</de> up to
           <de ID="de_18">the tanker car</de>
15.15  : _and_
15.16
: and send <de ID="de_19">it</de> back
           to Elmira

<link href="coref.xml#id(de_16)" type="ident">
  <anchor href="coref.xml#id(de_15)"/>
</link>
<link href="coref.xml#id(de_17)" type="ident">
  <anchor href="coref.xml#id(de_16)"/>
</link>
<link href="coref.xml#id(de_19)" type="ident">
  <anchor href="coref.xml#id(de_17)"/>
  <anchor href="coref.xml#id(de_18)"/>
</link>
```

As in the case of DRAMA, we recommend to mark up all ambiguities, and not annotate 'vague' or 'ambient' references.

5 References to Visual Situation

In multimodal dialogues it is possible to refer to objects which have not been previously introduced, but are 'accessible' by virtue of being part of the visual situation: examples are objects on the screen in the case of multimodal applications (Bruneseaux and Romary, 1998) and references to landmarks in the map in the MAPTASK corpus. The first proposed extension to the Core Scheme consists of a new set of elements introduced in order to annotate references to the visual situation.

We adopted for this purpose a variant of the <universe> mechanism used in the Bruneseaux-Romary scheme. The idea is to assign an ID to each object in the visual situation that can be referred to, and then represent references to these objects by means of the same <link> mechanism used for anaphoric relations in the Core Scheme. For each object in the visual situation, a <ue> element gets created; the <ue> elements are then grouped in a <universe> element, as follows:

(15)
```
coref.xml:

<universe ID="U1">
<ue ID="ue1"> Diamond mine </ue>
<ue ID="ue2"> Graveyard </ue>
<ue ID="ue3"> Fast running creek </ue>
<ue ID="ue4"> Fast flowing river </ue>
<ue ID="ue5"> Canoes </ue>
</universe>
FOLLOWER: Uh-huh. Curve round. To your right.
GIVER:    Uh-huh.
FOLLOWER: Right.... Right underneath
          <de ID="de50"> the diamond mine. </de>
           Where do I stop.
GIVER:    Well..... Do. Have you got
           de ID="de51"> a graveyard?</de>
           Sort of in the middle of the page?...
           On on a level to
           <de ID="de52"> the c--... er diamond
            mine. </de>

<link href="coref.xml#id(de50)" type="ident">
  <anchor href="coref.xml#id(ue1)"/>
  <anchor href="coref.xml#id(de_18)"/>
</link>
<link href="coref.xml#id(de51)" type="ident">
  <anchor href="coref.xml#id(ue2)"/>
  <anchor href="coref.xml#id(de_18)"/>
</link>
<link href="coref.xml#id(de52)" type="ident">
  <anchor href="coref.xml#id(ue3)"/>
  <anchor href="coref.xml#id(de_18)"/>
</link>
```

Having a single universe is sufficient in cases when there is a single case of objects, but not in domains like the MapTask, where the two participants to the conversation have slightly different maps. The <universe> mechanism has been designed to handle this type of situations, as well. In these cases, it is suggested that three universes be created: one with ID="COMMON" containing all objects shared between the visual situations, and then one universe for each conversational participant containing the elements known only to that participant. This will ensure that the shared elements receive a unique ID. <universe> elements have an optional MODIFIES attribute that can be used to encode the information that a given universe is an extension of another uni-

verse; e.g., in the case just discussed, the universe of each participant could be given a value for the MODIFIES="COMMON".

In (16b) we see how the situation in (16a) could be encoded. Three universes are defined; COMMON contains a gold mine, whereas the GIVER_UNIVERSE also contains a diamond mine, which isn't in the follower's universe. As a result, the follower mistakenly believes that the gold mine and the diamond mine are the same. This example also illustrates how these misunderstandings could be encoded by means of another optional extension to the link mechanism specified in the Core Scheme: the attribute WHO-BELIEVES, whose values would be identifiers for the two participants in the conversation (G and F in this case).

(16) a.

```
GIVER: Do_you have diamond_mine.
FOLLOWER: Yes I've got a gold_mine.
GIVER: Ah. S--.
FOLLOWER: ....
GIVER: You don't have diamond_mine though.
FOLLOWER: No. It's a gold_mine according to
       this one.
       Presumably that's the same.
GIVER: Well I've got a gold_mine as well
       you see. (MT)
```

b.

```
coref.xml:

<universe ID="common">
<ue ID="ue2"> gold mine </ue>
....
</universe>
<universe ID="GIVER_universe"
          modifies="common">
<ue ID="ue1"> diamond mine </ue>
...
</universe>
<universe ID="FOLLOWER_universe"
          modifies="common">
....
</universe>

GIVER: Do_you have
       <de ID="de_20"> diamond_mine. </de>
FOLLOWER: Yes I've got
       <de ID="de_21"> a gold_mine. </de>
GIVER: Ah. S--.
FOLLOWER: ....
GIVER: You don't have
       <de ID="de_22"> diamond_mine </de>
       though.
FOLLOWER: No.
       It's <de ID="de_23"> a gold_mine</de>
       according to this one.
       Presumably <de ID="de_24"> that's </de>
       the same.
GIVER: Well I've got
       <de ID="de_25"> a gold_mine </de>
       as well you see.

<link href="coref.xml#id(de_20)"  type="ident"
      who-believes="G">
   <anchor href="coref.xml#id(ue1)"/>
</link>
<link href="coref.xml#id(de_21)"  type="ident"
      who-believes="F" >
   <anchor href="coref.xml#id(ue2)"/>
</link>
<link href="coref.xml#id(de_21)"  type="ident"
      who-believes="F" >
   <anchor href="coref.xml#id(de_20)"/>
</link>
<link href="coref.xml#id(de_22)"  type="ident"
      who-believes="G">
   <anchor href="coref.xml#id(ue1)"/>
</link>
<link href="coref.xml#id(de_22)"  type="ident">
   <anchor href="coref.xml#id(de_20)"/>
</link>
<link href="coref.xml#id(de_23)"  type="ident"
      who-believes="F" >
   <anchor href="coref.xml#id(ue2)"/>
</link>
<link href="coref.xml#id(de_23)"  type="ident"
      who-believes="F" >
   <anchor href="coref.xml#id(de_22)"/>
</link>
<link href="coref.xml#id(de_24)"  type="ident"
      who-believes="F" >
   <anchor href="coref.xml#id(de_22)" />
</link>
```

We don't know of any reliability study for this type of references, but experience with MapTask suggests that it can be done reliably. We are currently doing a test of the reliability of this extension in two languages (Italian and English) and will report at the Workshop.

6 Marking non-nominal elements

Even if we only consider anaphoric relations involving nominal elements, there are at least two situations in which an annotator may wish to mark an anaphoric relation that also involves other types of constituents. The first is the case, already mentioned in Section 4, in which we have a relation that would fall for all purposes under the Core Scheme, except that the anaphoric element is either unexpressed or incorporated in the verb. The second situation are the cases of so-called DISCOURSE DEIXIS (Webber, 1991), when the antecedent of a nominal expression is an abstract object such as an event or proposition introduced in the discourse somewhat indirectly by sentences. (DRAMA allows for such relations to be marked.)

The second extension to the Core Scheme was developed to give annotators tools to mark these types of anaphoric relations. The solution we propose is to use the <seg> element introduced in the TEI to mark up arbitrary pieces of text; <seg> elements are given an ID which can then be used in <link> elements just like for other anaphoric relations. (The <seg> element could also be used to extend the meta scheme to cover anaphoric relations between non-nominal elements, such as VP ellipsis.)

6.1 Using SEG to mark up empty and incorporated constituents

In Italian, Spanish and many other languages, certain nominal constituents may not be realized; this is especially common for nominals in subject position. These nominals are present in annotations produced

by hand (e.g., in the Penn Treebank), but the parsers used for parsing spoken dialogues tend not to produce representations containing empty constituents in this case. In case these nominals are not represented in the base level, we recommend to mark the verb with a <seg> element, and then code the anaphoric relation as usual by means of <link> elements, as follows:

(17) coref.xml:

```
A: Dov'e' <de ID="de_157">Gianni?</de>
   [Where is Gianni?]
B: <seg type="pred" ID="seg_158 >e'
   andato a mangiare </seg>
   [_ went to have lunch]

<link href="coref.xml#id(seg_158)"
      type="ident">
   <anchor href="coref.xml#id(de_157)"/>
</link>
```

The reader will have noticed that this representation can only be used without loss of information when there is at most one empty elements; this is true for Italian, but not for Japanese or Portuguese. If more precision needed, the annotator should then define more specific identity relations also specifying which empty argument of the verb enters in the anaphoric relation: SUBJ-IDENT, OBJ-IDENT, etc. These relations could then used instead of IDENT to specify the value of the TYPE attribute of the <link> element.

A second case in which an argument is not realized by means of a nominal is in the case of incorporated clitics, such as *daselo* in (11). In this case, again, we recommend marking the verb by way of a <seg> element when the parser doesn't produce a morphologically decomposed representation, and then encoding the anaphoric relations in which the clitics are involved by means of either a single IDENT relation or by means of more fine-grained relations such as SUBJ-IDENT or OBJ-IDENT.

(18) coref.xml

```
Mira, te doy <de ID="de_167"> este libro </de>
Conoces a <de ID="de_168"> mi suegra?</de>
Pues <seg ID="seg_169"> daselo</seg> cuando
<de ID="de_170"> la </de> veas.

<link href="coref.xml#id(seg_169)"
      type="obj-ident">
  <anchor href="coref.xml#id(de_167)"/>
</link>
<link href="coref.xml#id(seg_169)"
      type="iobj-ident">
  <anchor href="coref.xml#id(de_168)"/>
</link>
```

Provided that the <seg> elements are identified during the first pass of markable identification, encoding this information should not be any harder than in the case of the Core Scheme. The real question for this type of annotation is which empty elements to annotate –e.g., in addition to 'small pro' elements such as those discussed above, the annotator may also decide to annotate 'big PRO' elements that according to some syntactic theories occupy the subject position of infinitival clauses.

6.2 Using SEG to mark the antecedents of discourse deixis

Abstract objects such as events, actions and propositions can all serve as antecedents of anaphoric expressions. We are not aware of any reliability results for this type of annotation, but the <seg> element can be used to identify the antecedents in this type of anaphora. If desired, the annotator could use a second attribute TYPE to specify the type of object introduced by the <seg> element; TYPE would have values EVENT, PROP and ACTION.

(19) a. *The 23-year-old had hit his head against another player during a game of Aussie-rules football.* McGlinn remembered nothing of the collision, but developed a headache and had several seizures. (BBC)

b.
```
<seg type="event" ID="seg_130">The 23-year-old
had hit his head against another player</seg>
during a game of Aussie-rules football.
McGlinn remembered nothing of
<de ID="de_131"> the collision </de>,
but developed a headache and had several
seizures.

<link href="coref.xml#id(de_131)"
      type="ident">
  <anchor href="coref.xml#id(seg_130)"/>
</link>
```

(20) a. Despite the latest negative results, *doctors are still convinced that Tamoxifen can prevent breast cancer.* This is because of the way it blocks the action of oestrogen, the female sex hormone that can make the breast cells of some women go out of control.

b.
```
Despite the latest negative results,
<seg type="prop" ID="seg_129"> doctors
are still convinced that
<de ID="de_131"> Tamoxifen </de> can
prevent breast cancer </seg>.
<de ID="de_130"> This </de> is because
of the way <de ID="132"> it </de>
blocks the action of oestrogen, the
female sex hormone that can make the
breast cells of some women go out of
control.

<link href="coref.xml#id(de_130)"
      type="ident">
  <anchor href="coref.xml#id(seg_129)"/>
</link>
```

(21) a.
```
GIVER: You're sort_of going past stone creek...
       but your line's curving up past the...
       flat rocks.
FOLLOWER: Right. Okay.
GIVER: AND THEN STARTING TO COME DOWN AGAIN.
```

```
FOLLOWER: Got THAT
```

b.
```
GIVER: You're sort_of going past stone creek...
       but your line's curving up past the...
       flat rocks.
FOLLOWER: Right. Okay.
GIVER: <seg ID="seg_135" type="action">And
       then starting to come down again.</seg>
FOLLOWER: Got <de ID="de_136"> that </de>.

<link href="coref.xml#id(de_136)"
      type="ident">
  <anchor href="coref.xml#id(seg_135)"/>
</link>
```

These examples also ilustrate some of the problems to be addressed when designing a reliable annotation scheme for this phenomenon: these include deciding what part of the text counts as antecedent as well as deciding which type of object the antecedent is (see, e.g., (21)).

7 Bridging References

DRAMA also allows annotators to encode certain types of BRIDGING REFERENCES (Clark, 1977): these are anaphoric expressions that denote objects that have not yet been introduced in the discourse, but that are related to an entity already introduced in the text by relations other than identity. An example is *the indicators* in:

(22) John has bought *a new car. The indicators* use the latest laser technology.

We are able to interpret the description *the indicators* because we know that indicators are parts of cars. The set of relations that may hold between a bridging reference and its 'antecedent' or 'anchor' is rather wide; an extensive survey of the existing classifications can be found in (Vieira, 1998).

The Extended Relations Scheme is designed for those who wish to mark up this more general anaphoric relations. It uses the same elements as the Core Scheme, but more values are allowed for the TYPE attribute of the `<link>` element besides simple IDENT. The set of relations allowed by the scheme derives from the analysis of Vieira and includes most of the bridging relations in DRAMA (MEMBER, SUBSET, PART, CAUSE, POSS and ARG). For example, we see in (23) how the elements of the Extended Relations Scheme can be used to encode a subset relation between *les modeles de fusees* and *les fusees qui ont bien vole'*.

(23) a.
```
F: Alors donc / vous avez / ici /
   LES MODELES DE FUSEES /
M: Oui
F: Et vous allez essayer de vous
   mettre d'accord sur un classement
   /hein classer
   LES FUSEES QUI ONT BIEN VOLE' ou
   QUI ONT MOINS BIEN VOLE'
```

b.
```
F: Alors donc / vous avez / ici /
   <de ID="de_88"> les mode'les de fuse'es </de>
M: Oui
F: Et vous allez essayer de vous mettre d'accord
   sur un classement /hein classer
   <de ID="de_89"> les fuse'es qui ont
   bien vole' </de>
   ou <de ID="de_90"> qui ont
   moins bien vole' </de>

<link href="coref.xml#id(de_89)">
  <anchor href="coref.xml#id(de_88)"
          type="subset " />
</link>
<link href="coref.xml#id(de_90)"
      type="subset " >
  <anchor href="coref.xml#id(de_88)"/>
</link>
```

As the poor reliability scores which have been obtained by (Poesio and Vieira, 1998) for this kind of scheme indicate, once one moves beyond the ident relation, it can be difficult to decide how to classify the link between two elements. We tried to alleviate this problem by adopting the TEI technique of specifying 'subtypes' of links: in those cases in which it may be difficult to identify precisely the type of relation that exists between two entities, we introduced a more general relation to be used as type of a link, as well as more specific relations to be used as values of the SUBTYPE attribute in those cases in which this additional specification is possible. We used this technique for two types of relations: possession relations (which include generic attribution, true possession and part as subtypes) and event relations (which include relations such as cause and 'role' as subparts). The following example illustrates how type and subtype attributes can be used to encode possession relations at the desired level of precision, as well as why sometimes it may be difficult to decide which relation holds between two discourse entities.

(24) a. French boss Aime Jacquet praised *his team's application* (BBC)

b.
```
<de ID="de_91"> French boss
Aime Jacquet </de> praised
<de ID="de_92">
<de ID="de_93">
 <de ID="de_94"> his </de>
team's </de> application. </de>

<link href="coref.xml#id(de_94)"
            type="ident" >
<anchor href="coref.xml#id(de_91)"/>
</link>
<link href="coref.xml#id(de_93)"
            type="poss " subtype="sposs" >
<anchor href="coref.xml#id(de_94)" />
</link>
<link href="coref.xml#id(de_92)"
             type="poss " subtype="attr " >
<anchor href="coref.xml#id(de_93)"/>
</link>
```

In the documentation we specify additional relations and further distinctions that an annotator may wish to make, including ways to annotate the function-value relations discussed in MUCCS.

The basic problem to be solved when trying to do this type of annotation is to come up with instructions that will ensure that annotators recognize bridging references. As a preliminary proposal, we suggest that annotators try first to identify an antecedent which is identical with the anaphor; if that fails, they should try first to find a discourse entity with which the anaphor stands in one of the set relations, then one with which it stands in one of the generalized possession relations.

8 State of the Proposal; Further Work

So far, we have used to scheme to annotate a TRAINS dialogue, a MAPTASK dialogue, and a dialogue from the microfusees corpus collected by LORIA. We are currently running a reliability study of the extension dealing with references to the visual situation, while waiting for the preliminary release of the MATE workbench to study the more complex features of the scheme. This will also involve trying to extract a coding book from the manual by fixing up some parameters. As the preliminary release of the MATE workbench is planned for May, we may be able to report some results already at the ACL meeting.

Acknowledgments

The MATE project is supported by European Union, Telematics LE4-8370. Massimo Poesio is supported by an EPSRC Advanced Research Fellowship.

References

J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. 1997. Dialogue acts in VERBMOBIL-2. Verbmobil Report 204, DFKI.

J. Allen and M. Core. 1997. DAMSL: Dialogue act markup in several layers. Draft contribution for the Discourse Resource Initiative, October.

F. Bruneseaux and L. Romary. 1998. Documents préparatoires pour le codage de dialogues multimodaux suivant les directives de la TEI. Available at `http://www.loria.fr/~romary/Documents/index.html`.

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32.

H. H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press.

Discourse Resource Initiative. 1997. Standards for dialogue coding in natural language processing. Report no. 167, Dagstuhl-Seminar.

S. Fligelstone. 1992. Developing a scheme for annotating text to show anaphoric relations. In G. Leitner, editor, *New directions in English language corpora*, pages 153–170. de Gruyter.

U. Heid and A. Mengel. 1999. A query language for research in phonetics. In *ICPhS*, San Francisco.

I. Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts at Amherst, 1982.

L. Hirschman. 1997. MUC-7 coreference task definition, version 3.0. Available at `http://www.muc.sail.com/proceedings/co_task.html`.

H. Kamp. 1981. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, *Formal Methods in the Study of Language*, 277-322. Mathematisch Centrum.

R. Passonneau. 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.

V. Pirrelli and C. Soria. 1999. Morpho-syntax annotation scheme. Available on the Web, February.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June.

R. Quirk and S. Greenbaum. 1973. *A University Grammar of English*. Longman.

K. van Deemter and R. Kibble. 1999. What is coreference, and what should coreference annotation be? In *Proc. of the ACL Workshop on Coreference*, Maryland.

R. Vieira. 1998. *Definite Description Resolution in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Centre for Cognitive Science.

B. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.