The VENEX corpus of anaphora and deixis in spoken and written Italian

Massimo Poesio[†], Rodolfo Delmonte[§], Antonella Bristot[§], Luminita Chiran[§], and Sara Tonelli[§]

[†]University of Essex
[§] Universitá di Venezia

Abstract

The VENEX corpus is a corpus of Italian annotated with information about anaphora and deixis, created in a joint project between the Università di Venezia and the University of Essex. The corpus includes both texts (articles from a financial newspaper) and dialogues (an Italian version of the MapTask corpus). The annotation scheme is an almost complete implementation of the scheme proposed in MATE, and the markup scheme is the simplified form of standoff adopted in the MMAX annotation tool.

1. INTRODUCTION

The MATE 'meta-scheme' scheme for anaphora annotation (Poesio et al., 1999) is one of the annotation schemes developed as part of the MATE project (McKelvie et al., 2001) The MATE proposals have served as the basis for a number of annotation projects, especially the development of the GNOME corpus (Poesio, 2000; Poesio et al., 2004b), as well as to the development of tools for anaphora annotation, such as MMAX (Müller and Strube, 2003). However, not all aspects of the original recommendations have been tested. Aspects never tried before include the recommendations for marking anaphoric reference to landmarks in MapTask-style situations (Anderson et al., 1991), and for dealing with anaphoric elements that are either unrealized or incorporated in other elements, e.g., in Italian. From the point of view of annotation technology, the GNOME annotation did not adopt one of the central aspects of the MATE proposals, standoff; in addition, in MATE it was recommended to do anaphoric annotation off the output of a parser, whereas in GNOME markables were identified by hand.

All of these aspects of the proposals have been tested during the creation of the VENEX corpus, a joint project between the Università di Venezia and the University of Essex. The corpus includes both texts and dialogues. The annotation scheme is an almost complete implementation of the scheme proposed in MATE, and the markup scheme is the simplified form of standoff adopted in the MMAX annotation tool. The work on VENEX has led to a number of developments of the original proposals, as well as to a re-examination of a number of their aspects. In this paper, we consider some of the issues raised by this work, and discuss they were addressed.

2. THE MATE ANNOTATION SCHEME FOR 'COREFERENCE'

We summarize in this section the most distinctive features of the MATE scheme. A complete description of the MATE scheme is available from the MATE project pages at `http://mate.nis.sdu.dk/`.

2.1. The Markup Scheme

The core aspect of the MATE proposals is the scheme for marking up anaphoric relations in XML. As in the MUC scheme (MUCCS) (Hirschman, 1998), it is assumed that annotation of anaphora is best separated in two steps: first the *markables* (the text constituent that realize semantic objects that may enter in anaphoric relations) are agreed upon, then anaphoric relations between them are marked. The main difference from the MUC scheme is that whereas in MUCCS anaphoric relations are annotated using attributes on the markables, in the MATE scheme–following the recommendations of the Text Encoding Initiative (Burnard and Sperberg-McQueen, 2002), and of Bruneseaux and Romary (1998)–the distinction between these two steps of annotation is mirrored by a distinction between two XML elements: ⟨`de`⟩, used to indicate the markables, and ⟨`link`⟩, used to mark information about these relations. ⟨`link`⟩ elements are *structured* elements, containing one or more ⟨`anchor`⟩ element. The ⟨`link`⟩ element specifies the anaphoric expression (using XML's `HREF` mechanism) and the relation between the anaphoric expression and its antecedent; whereas the ⟨`anchor`⟩ element specifies the antecedent.

(1) `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>
```

The design of the MATE workbench was strongly inspired by the concept of *standoff* annotation introduced for the MapTask. The main principle of standoff annotation is that each level of annotation–for example, syntactic annotation, dialogue act annotation, and anaphoric annotation–should be stored independently; in this way, annotators working on one level need not be concerned about the other levels of annotation, and can start immediately without having to wait for other annotation tasks to be completed. The separate levels of annotation are synchronized via a *base file*, to which the separate levels point using the HREF mechanism of XML. For the coreference scheme, as well, it was proposed that ⟨link⟩ elements should be kept in separate files pointing at the file in which the ⟨de⟩ elements were indicated.

2.2. Instantiations of the Meta-Scheme

One of the most important assumptions behind the design of the MATE proposals for anaphoric annotation is the belief that given the variety of phenomena that go under the name of anaphora, and the variety of possible applications, there can be no such thing as a general-purpose scheme for anaphoric annotation. Instead, it was shown how the basic mechanisms discussed above could be used to implement different types of anaphoric annotation, including some of the most popular schemes for 'coreference annotation,' such as MUCCS, Passonneau's DRAMA scheme (1997) , and the scheme used for annotation of references to landmarks in the MapTask corpus.

The Core Scheme In the most basic type of coreference scheme, such as MUCCS, only anaphoric relations between NPs are considered, and only identity relations. Schemes of this type can be implemented by allowing for only one anaphoric relation, IDENT. The remaining differences between the schemes have then mostly to do with the instructions to annotators–for example, which types of anaphoric relations to be considered as cases of 'identity' (see (van Deemter and Kibble, 2000) for some problems with the choices made in MUC).

Extended Relations DRAMA extends such schemes with ways of annotating associative relations. References of this type can be annotated in the MATE markup scheme using additional relations, as in (2).

(2) 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>
```

It was pointed out, however, that the results of Poesio and Vieira (1998) indicated that this type of annotation could be highly unreliable.

References to the Visual Situation A special ⟨universe⟩ element was suggested for MapTask-style annotations of references to visible objects. The ⟨universe⟩ element containing one ⟨ue⟩ element for each object in the visual scene; including such elements in an annotation makes it possible to use ⟨link⟩ elements to annotate references to such objects. Cases in which the participants to a conversation have different visual situations, as in the MapTask dialogues, can be handled by having separate universes, one for each participant to the conversation. In addition, a WHO-BELIEVES attribute of ⟨link⟩ elements was proposed to represent situations in which only one participant believes that a particular anaphoric relation holds, as in the following example, where only Follower believes that the 'gold mine' refers to the same object as the 'diamond mine'.

(3) 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>
```

2.3. Instructions for Identifying Markables

One of the novel aspects of the MATE instructions was the concern for markable identification in languages other than English. One such issue was how to deal with incorporated clitics and empty subjects. The suggestion contained in the MATE guidelines was to use a separate element, ⟨seg⟩, to turn verbs into non-nominal markables, as in the following example:

(4)
```
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 ⟨seg⟩ element was also meant to be used in more ambitious schemes as general mechanism for specifying non-nominal markables –e.g., to indicate the antecedents of discourse deixis, or for ellipsis.[1]

3. AN INSTANCE OF THE MATE META-SCHEME: GNOME

Ideas from the MATE 'scheme' have been adopted and tested both in annotation projects, such as the development of the GNOME corpus, and by the developers of annotation tools. The GNOME corpus was developed to study discourse properties claimed to affect the way discourse entities are realized, including definiteness (Poesio, 2004) and salience, particularly as formalized in Centering theory (Poesio et al., 2004b) and Grosz and Sidner's theory of the attentional state (Poesio and Di Eugenio, 2001). These studies were in part motivated by work on natural language generation, and fed into a series of papers studying sentence planning (Poesio, 2000; Henschel et al., 2000; Cheng et al., 2001) and text planning (Karamanis, 2003; Kibble and Power, 2003). The corpus is also being used to study anaphora resolution (Poesio and Alexandrov-Kabadjov, 2004), with a special focus on the resolution of bridging references (Poesio, 2003; Poesio et al., 2004a). This work led both to the development of a detailed coding manual for the parts of the MATE proposals incorporated in the GNOME scheme, and to further developments. In this section we briefly discuss how the MATE scheme was used and further developed in GNOME, particulary as far as the annotation of bridging references and deixis is concerned. For further details about the GNOME corpus and for the complete annotation manual, see `http://hcrc.ed.ac.uk/~gnome`.

3.1. Genres

The GNOME corpus includes texts from three domains. The museum subcorpus consists of descriptions of museum objects, generally with an associated picture, and brief texts about the artists that produced them. The pharmaceutical subcorpus is a selection of leaflets providing the patients with legally mandatory information about their medicine.

3.2. Markup scheme

Several layers of information were annotated, including layout in the case of text and rhetorical structure in the case of tutorial dialogues, sentences and potential utterances, noun phrases, a variety of attributes of the objects denoted by noun phrases,[2] and anaphoric relation. We concentrate here on anaphoric information, and refer the reader to the manual available from (`http://hcrc.ed.ac.uk/~gnome`) for the other types of annotation.

The parts of the GNOME annotation scheme that have to do with anaphora implement are based on

[1] A second range of issues considered in the MATE scheme had to do with dialogue phenomena, such as non-contiguous elements; we will not consider these issues here.

[2] E.g., whether an NP denoted generically or not; whether it denoted an animate or inanimate entity, as well as other ontological properties; and whether it denoted a discourse entity, a quantifier, or a predicate. In the case of a discourse entity, we also annotated whether it denoted an atom, a set, or a mass term; and whether it denoted uniquely.

the 'Core' and the 'Extended Relations' instantiations of the MATE meta-scheme. The markup scheme for markables and anaphoric relations adopted in GNOME follows very closely that proposed in MATE, except that the ⟨de⟩ element was renamed ⟨ne⟩, and the ⟨link⟩ element was renamed ⟨ante⟩. More substantial differences are the decision not to use standoff, and the introduction of new elements necessary for the study of salience, such as elements that could be used to investigate the notion of UTTERANCE used in Centering (Poesio et al., 2004b).

Although standoff is a clear improvement over including all annotation levels in a single file, our own experiences during the creation of the GNOME corpus being further proof of this, it's only really possible when tools are available both to create the annotation and–crucially–later to 'knit back' the separate levels when needed. As neither the MATE workbench nor any other tools based on standoff were available by the time the GNOME annotation started,[3] in GNOME we didn't use standoff, but integrated all levels of annotation in one file; an Emacs minor mode extending SGML-mode, GNOME-mode, was developed.[4]

The main new aspect of the markup scheme, especially as far as our studies of salience were concerned, is the inclusion of elements used to annotate potential utterances in the sense of Centering (Grosz et al., 1995). In order not to prejudge the answer to the question of which text constituents are best viewed as utterances, we used a 'generic' element called ⟨unit⟩ to mark up finite and non-finite clauses, but also parentheticals and elements of bulleted lists.

3.3. Bridging References and Deixis

Apart from the relation of identity, in GNOME we were concerned with bridging references and deictic reference, hence the annotation scheme incorporated aspects of the 'Extended Relations' and the 'MapTask' instantiations of the MATE meta-scheme.

One of our aims was to continue the work on bridging references annotation and interpretation in (Poesio and Vieira, 1998; Vieira and Poesio, 2000), which showed that marking up bridging references is quite hard. In addition, work such as (Sidner, 1979; Strube and Hahn, 1999) suggested that indirect realization can play a crucial role in maintaining the CB. After testing a few types of associative reference (Hawkins, 1978), we decided to annotate only three non-identity relations, as well as identity. These relations are a subset of those proposed in the 'extended relations' version of the MATE scheme: set membership (ELEMENT), subset (SUBSET), and 'generalized possession' (POSS), which includes both part-of relations and ownership relations.

In our preliminary attempts at annotating deictic references we used a technique similar to the 'Universe' scheme developed in MATE. However, we quickly realized that, first of all, 'real' pictures cannot be decomposed into 'objects' as easily as the maps used in the MapTask, hence asking the annotators to identify specific objects as the referents of deictic references was quite hard. Secondly, that none of the studies we intended to carry out actually required this identification; all that was needed –e.g., to study the use of demonstratives–was to know whether a reference was deictic or not. As a result, we used a boolean DEICTIC attribute.

3.4. Coder manual

Perhaps the most important aspects of the GNOME annotation are the development of detailed instructions for annotators (see http://hcrc.ed.ac.uk/~gnome) and the reliability experiments testing several aspects of the scheme, particularly bridging references.

The identification of sentences, units and markables was done entirely by hand, without encountering particular problems. All attributes of sentences, ⟨unit⟩s and ⟨ne⟩s in the final version of the scheme, including DEIX, can be annotated reliably. In order to achieve reliability on anaphoric annotation, the range of anaphoric phenomena considered was restricted in many ways. Apart from marking a limited number of associative relations, the annotators only marked relations between objects realized by noun phrases and not, for example, anaphoric references to actions, events or propositions implicitly introduced by clauses or sentences. We also gave strict instructions to our annotators concerning how much to mark. We found a rather good agreement on identity relations. In our most recent analysis (two annotators looking at the anaphoric relations between 200 NPs) we observed no real disagreements; 79.4% of these relations were marked up by both annotators; 12.8% by only one of them; and in 7.7% of the cases, one of the annotators marked up a closer antecedent than the other. Concerning associative references, limiting the relations did limit the disagreements among annotators (only 4.8% of the relations are actually marked differently) but only 22% of bridging references were marked in the same way by both annotators; 73.17% of rela-

[3] In the end lack of time prevented the inclusion of a tool for anaphoric annotation in the released workbench.

[4] GNOME-mode provides some support for introducing new elements, marking regions, and attribute editing, as well as anaphoric annotation.

tions are marked by only one or the other annotator. Reaching agreement on this information involved several discussions between annotators and more than one pass over the corpus (Poesio, 2000).

4. VENEX: GOALS, MARKUP, AND ANNOTATION SCHEME

The general goals of the VENEX annotation were to produce a resource that could be used to study both pronominal and full NPanaphora in Italian, both from an anaphora resolution and from an anaphora generation perspective. More specific goals included conducting for Italian a study of the effect of Centering on pronominal anaphora analogous to (Poesio et al., 2004b), also taking into account the work of (Di Eugenio, 1998); and a study of definite description use like (Poesio and Vieira, 1998), but looking also at deictic references to the visual situation in dialogues. The ultimate goal is to use the corpus to test anaphora resolution algorithms for Italian.

From a linguistic point of view, the more substantial difference between the VENEX annotation and the annotation effort in GNOME is the inclusion of dialogue data and the need to consider a variety of forms of anaphoric reference not present in English. This made it necessary to consider issues addressed in the MATE guidelines but not relevant for GNOME. The VENEX annotation scheme incorporates aspects of all three instantiations of the MATE meta-scheme– core, extended relations, and references to the visual situation– as well as the suggestions for dealing with clitics, zero anaphora, and misunderstandings.

The VENEX annotation also goes beyond GNOME in that more modern annotation technology is used, in two respects: markables are identified automatically as far a possible, and data are stored in a standoff format, using a modern annotation tool (MMAX).

4.1. The Data

The VENEX annotation effort builds on the results of two separate corpus-annotation initiatives: SI-TAL (Montemagni, 2000), concerned with the creation of a corpus of written Italian text from financial newspapers (*Il Sole 24 Ore*) comparable to the Wall Street Journal corpus; and IPAR (Bristot et al., 2000), a continuation of the previous projects API and AVIP projects, whose result was a collection of spoken task-oriented dialogues of speakers performing the Map-Task (Anderson et al., 1991). The VENEX corpus consists of 30 SI-TAL newspaper articles and 6 IPAR dialogues.

4.2. The Annotation Tool

As the choice of the annotation tool played an important role in the design of the markup scheme, we will discuss this first. Having observed the difficulties annotators had during GNOME, we wanted to use a proper annotation tool, possibly one based on standoff technology. No annotation tool implementing the MATE or GNOME schemes as described exists; but in the years after the development of the MATE guidelines tools supporting XML standoff annotation for coreference have appeared, including MMAX from EML (Müller and Strube, 2003) and the Annotator from ILSP. Although the format used for storing anaphoric information by these tools is not entirely satisfactory, the files they produce can be easily converted into MATE format.

The tool used in VENEX, MMAX is based on a simplified standoff format without `href` references to the base file. Three main files are maintained for each annotated file in the corpus: a base file containing the words, a file specifying how the text is broken up (into paragraphs and sentences in the case of written text, into turns in the case of dialogues), and a file identifying markables. A special `.anno` file records the names of the three files, which have to be kept in the same directory. The word level and markable level files for one of the files in the VENEX corpus, `napoli-05`, are shown in Fig. 1.

Word files contain one ⟨`word`⟩ element per token, with a unique `ID`. Text files for written text consist of a ⟨`text`⟩ element containing one or more ⟨`paragraph`⟩ elements, in turn containing one or more ⟨`sentence`⟩ elements with a unique ⟨`id`⟩ and a ⟨`span`⟩ indicating the words belonging to the sentence. Turn files for dialogue consist a ⟨`turns`⟩ element including one or more ⟨`turn`⟩ elements which, in addition to ⟨`id`⟩ and ⟨`span`⟩, contain an optional ⟨`speaker`⟩ attribute. Examples of both types of files are given below.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE words SYSTEM "text.dtd">
<text id="1">
<paragraph id="paragraph_1">
<sentence id="sentence_1" span="word_1..word_13"/>
</paragraph>
....
</text>

<?xml version="1.0" encoding="iso-8859-1"?>
<turns>
<turn id="turn0" span="word_1..word_19"
      speaker="g001"/>
.....
</turns>
```

The most important file for our purposes is the markable file, which contains a ⟨`markables`⟩ element

containing one or more ⟨markable⟩ elements. In addition to ⟨id⟩ and ⟨span⟩ elements, markables have two special attributes used to store anaphoric information: the MEMBER attribute, used to indicate membership in a coreference chain (a coreference equivalence class), and the POINTER attribute, used to mark up to one associative anaphoric relation for each anaphoric expression. Any number of additional attributes for ⟨markable⟩ elements can be specified by users via a COREFERENCE SCHEME.

4.3. Adapting the MATE Markup Scheme For Use with MMAX

The MATE / GNOME markup scheme had to be adapted in a number of ways to be usable with MMAX. The first issue was that MMAX doesn't support ⟨link⟩ elements–anaphoric information is stored with markables–so in the VENEX annotation we had to use markable attributes to represent information that in MATE and GNOME-style markup schemes would have been encoded as part of the ⟨link⟩ elements. We used a separate attribute to specify the type of associative relation expressed by POINTER attribute, and a SPACE attribute to encode the information stored in the WHO-BELIEVES attribute of links (see below). In addition, only one MEMBER and POINTER attributes can be specified for each markable. This latter limitation wasn't much of a problem, given that the annotation instructions used in VENEX are derived from those developed for GNOME and also attempt to limit annotators to mark at most one identity and one bridging relation for each anaphoric expression. The separation of attributes of links proved, however, a problem, as annotators often forget to annotate one or the other.

A second problem is that the version of MMAX we used (0.94) only allows for one type of markable, meaning that ⟨unit⟩ elements could not be annotated, and instead of using separate ⟨ne⟩ and ⟨seg⟩ elements for nominal and non-nominal markables, a single markable had to be used (see below).[5] A boolean attribute SOGGETTO_VUOTO was used; SOGGETTO_VUOTO = "none" for normal ⟨ne⟩ markables, whereas SOGGETTO_VUOTO = "true" for ⟨markable⟩ elements that do the job of ⟨seg⟩ elements, marking verbal elements that contain incorporated clitics or when the subject is dropped.

The complete list of user-defined markable attributes currently being annotated is in Table 1.

Two attributes, ANAPHORIC_SPACE and BRIDGING_SPACE, are used to realize the WHO-BELIEVES attribute of the ⟨link⟩ element in the MATE scheme. FOr the moment, deictic function has been implemented with a simple boolean attribute, as done in GNOME.

[5]The version of MMAX currently being developed will allow for multiple markables.

Attribute	Brief explanation
np_form	Type of NP (the-np, etc.)
soggetto_vuoto	Boolean: "true" for ⟨seg⟩ elements
is_anaphoric	Whether the head predicate of the NP has already been used to describe the entity
anaphoric_space	Which discourse models contain the identity relation
is_bridging	Type of associative relation (part-of, element, set, attribute)
bridging_space	Which discourse models contain the bridging relation
function	Whether the NP refers deictically

Table 1: Markable attributes

5. VENEX: ANNOTATION METHODOLOGY

5.1. Parsing

Whereas in the GNOME annotation annotators had to add markables by hand, the VENEX annotation is more alike the type of annotation originally envisaged in MATE, whereby markables have largely been identified automatically, and then corrected by hand.

Both the written and spoken corpora were tokenized, POS-tagged and parsed using a suite of tools developed at the Universitá di Venezia, including the IMMORTALE POS-tagger (Delmonte and Piana, 1999) and the GETA-RUN parser (Delmonte, 2002). A morphological analyzer extracts a number of features, including agreement features; GETA-RUN builds a complete constituent and functional structure. The output of the parser is then corrected semi-automatically using separate annotation tools.

A series of scripts converts the (different) formats used in SI-TAL and in IPAR into the MMAX format, identifies the markables, and automatically computes the type of the NPs.

5.2. Annotation procedure

A new and detailed coding manual was produced for the project. The instructions are mostly derived from those developed for GNOME; we briefly summarize here the main differences.

Markable correction The initial list of markables has to be corrected by hand. Although many details of this correction task are likely to be specific to the particular parser used, a few problems will probably have to be considered by other similar annotations as well.

One obvious problem are incorporated clitics and empty subjects; the parser cannot identify those, so in these cases, the annotators have to mark an element of the verbal complex as a ⟨markable⟩. (We saw above that whereas in the MATE scheme a separate the ⟨seg⟩ element for this purpose, in VENEX the same type of markable is used in both cases, but with an attribute specifying the type of markable.) The parser also misses a few nominal markables, especially pronouns in possessive position and nominals in certain types of coordinated constructions.

Misunderstandings The MapTask part of the VENEX corpus contains numerous examples like (3), where the differences between Giver and Follower map lead to one participant believing that two objects are anaphorically related, while the other participant either is not aware of this or doesn't believe this to be the case. We found that after a few iterations of training, our annotators were able to handle these cases properly (a more formal evaluation is underway; we hope to report the results at the meeting). Again, the only problems were caused by the fact that these attributes had to be added to markables, which sometimes led to annotators forgetting to set them. (This was only required in case the default, that an anaphoric relation was in the common ground of both participants, didn't hold.)

5.3. State of the Annotation

The entire corpus has been annotated; we are currently running new reliability studies, and will then revise the annotation. We expect the work to be completed in the Summer, and the corpus to be made available at the end of the year or early next year.

6. DISCUSSION

We briefly discussed the annotation scheme and methods used to create the VENEX corpus. This experience has prompted a reconsideration of the original MATE recommendations for anaphoric annotation. We discuss a few issues directly related to the question of using XML for this type of annotation.

LINK elements Our experience with VENEX suggests that having a separate ⟨link⟩ element would be very useful; in fact, two of the most beneficial aspects that would derive from this we had not originally considered. First of all, separate ⟨link⟩ elements can be used to mark general semantic relations, not just anaphoric relations (for more complex types of semantic annotation). Secondly, and perhaps most importantly, grouping all attributes relevant to links into a single element makes it harder for annotators to forget to fill in aspects of the annotation. Unfortunately, at the moment there is no tool that can be used to create this type of annotation directly.

Anaphoric relations One aspect of the markup scheme that needs revision is the placement of the relation. One problem we observed in GNOME is that often the ambiguity is not simply between two possible antecedents each of which stands in the same relation to the anaphoric expression, but between two antecedents which stand in different relations. In the pharmaceutical texts, for example, it is often unclear whether a particular mention of the medicine under consideration refers to the generic product, or to the particular instance that the user has in their hands. In this case, we would want annotators to mark the anaphoric expression as IDENT with one object, and ELEMENT of the other (ELEMENT is also used in GNOME for relations between instances and types), as follows, but this is not possible in either the original MATE scheme or in the GNOME markup scheme:

(5)
```
<ante current="ne1">
<anchor ID="ne2" rel="ident">
<anchor ID="ne3" rel="element">
</ante>
```

Ambiguity Offering annotators the opportunity to annotate anaphoric ambiguity is essential, especially for annotations used to study linguistic phenomena, but raises serious theoretical and practical problems. A coreference chain containing such links becomes a coreference (directed) *graph*, in which each of the paths across the graph is a potential interpretation. While having multiple paths is not a problem as far as evaluating the results of an anaphoric resolver (any path in the graph counts as a valid solution), it is a serious problems both for scripts attempting to ensure consistency (e.g., that all references to the same object are marked as either generic or non-generic–this is of course impossible when one of the possible antecedents is generic while the other isn't) as well for annotation tools (the problem is of course worsened when the tool only uses a single attribute to indicate membership in a coreference chain).

Revision A second difficult problem is caused by cases, common in the MapTask dialogues, in which after a while a participant realizes that their previous belief that an object was identical to another object is mistaken. In these cases, the participant is arguably revising their previous beliefs; it is not clear then what should be done with the annotation of the original anaphoric information.

7. References

A. H. Anderson, M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAl-

lister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

A. Bristot, L. Chiran, and R. Delmonte. 2000. Verso un'annotazione XML di dialoghi spontanei per l'analisi sintattico-semantica. In *XI Giornate di Studio GFS, Multimodalita' e Multimedialit nella comunicazione*, pages 42–50, Padova.

F. Bruneseaux and L. Romary. 1998. Documents préparatoires pour le codage de dialogues multimodaux suivant les directives de la TEI.

L. Burnard and C. M. Sperberg-McQueen. 2002. TEI lite: An introduction to text encoding for interchange. http://www.tei-c.org/Lite.

H. Cheng, M. Poesio, R. Henschel, and C. Mellish. 2001. Corpus-based NP modifier generation. In *Proc. of the Second NAACL*, Pittsburgh.

R. Delmonte and E. Piana. 1999. Tag disambiguation in Italian. In *Proc. of the ATALA Workshop on Treebanks*, pages 41–49.

R. Delmonte. 2002. GETARUN: A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG. In *Proc. of LFG-02*, pages 130–153.

B. Di Eugenio. 1998. Centering in italian. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, pages 115–138. Oxford.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225.

J. A. Hawkins. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

R. Henschel, H. Cheng, and M. Poesio. 2000. Pronominalization revisited. In *Proc. of 18th COLING*, Saarbruecken, August.

L. Hirschman. 1998. MUC-7 coreference task definition, version 3.0. In N. Chinchor, editor, *In Proc. of the 7th Message Understanding Conference*.

N. Karamanis. 2003. *Entity coherence for descriptive text structuring*. Ph.D. thesis, Univ. of Edinburgh.

R. Kibble and R. Power. 2003. Optimising referential coherence in text generation. Submitted.

D. McKelvie, A. Isard, A. Mengel, M. B. Moeller, M. Grosse, and M. Klein. 2001. The MATE workbench - an annotation tool for XML corpora. *Speech Communication*, 33(1-2):97–112.

L. Montemagni. 2000. The italian syntactic-semantic treebank: Architecture, annotation, tools and evaluation. In *Proc. of LINC*, pages 18–27.

C. Müller and M. Strube. 2003. Multi-level annotation in MMAX. In *Proc. of the 4th SIGDIAL*, pages 198–207.

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

M. Poesio and M. Alexandrov-Kabadjov. 2004. A general-purpose, off the shelf anaphoric resolver. In *Proc. of LREC*, Lisbon, May.

M. Poesio and B. Di Eugenio. 2001. Discourse structure and anaphoric accessibility. In Ivana Kruijff-Korbayová and Mark Steedman, editors, *Proc. of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

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

M. Poesio, F. Bruneseaux, and L. Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker, editor, *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004a. Learning to solve bridging references. Submitted.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. M. Hitzeman. 2004b. Centering: A parametric theory and its instantiations. *Computational Linguistics*. To appear.

M. Poesio. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proc. of the 2nd LREC*, pages 211–218, Athens, May.

M. Poesio. 2003. Associative descriptions and salience. In *Proc. of the EACL Workshop on Computational Treatments of Anaphora*, Budapest.

M. Poesio. 2004. An empirical investigation of definiteness. In S. Kepser, editor, *Proc. of the International Conference on Linguistic Evidence*, Tübingen, January. University of Tübingen, SFB 441.

C. L. Sidner. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

M. Strube and U. Hahn. 1999. Functional centering. *Computational Linguistics*, 25(3):309–344.

K. van Deemter and R. Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637. Squib.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4), December.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE words SYSTEM "words.dtd">
<words>
<word id="word_1">G001</word>
<word id="word_2">Sara</word>
<word id="word_3">,</word>
<word id="word_4">allora</word>
<word id="word_5">c</word>
<word id="word_6">hai</word>
<word id="word_7">sulla</word>
<word id="word_8">tua</word>
<word id="word_9">sinistra</word>
<word id="word_10">-</word>
<word id="word_11">una</word>
<word id="word_12">figura</word>
....
</words>
```

```
<?xml version="1.0"?>
<markables>
<markable id="markable_1" span="word_25" type="verbalspec" np_form="clitic"/>
...
<markable id="markable_297" span="word_11..word_18" type="verbalspec"
np_form="none" function="deictic" soggetto_vuoto="none" member="set_1"
is_anaphoric="none" is_bridging="none" anaphoric_space="ana_both"
bridging_space="bridging_both"/>
```

Figure 1: Standoff annotation in MMAX: Words and Markables