# Discourse Annotation and Semantic Annotation in the GNOME Corpus

**Massimo Poesio**
University of Essex,
Department of Computer Science and Centre for Cognitive Science,
United Kingdom

## Abstract

The GNOME corpus was created to study the discourse and semantic properties of discourse entities that affect their realization and interpretation, and particularly salience. We discuss what information was annotated and the methods we followed.

## 1 Introduction

The GNOME corpus was created to study the aspects of discourse that appear to affect generation, especially salience (Pearson et al., 2000; Poesio and Di Eugenio, 2001; Poesio and Nissim, 2001; Poesio et al., 2004b). Particular attention was paid to the factors affecting the generation of pronouns (Pearson et al., 2000; Henschel et al., 2000), demonstratives (Poesio and Nygren-Modjeska, To appear) possessives (Poesio and Nissim, 2001) and definites in general (Poesio, 2004a). These results, and the annotated corpus, were used in the development of both symbolic and statistical natural language generation algorithms for sentence planning (Poesio, 2000a; Henschel et al., 2000; Cheng et al., 2001), aggregation (Cheng, 2001) and text planning (Karamanis, 2003). The empirical side of the project involved both psychological experiments and corpus annotation, based on a scheme based on the MATE proposals, as well as on a detailed annotation manual (Poesio, 2000b), the reliability of whose instructions was tested by extensive experiments (Poesio, 2000a). More recently, the corpus has also been used to develop and evaluate anaphora resolution systems, with a special focus on the resolution of bridging references (Poesio, 2003; Poesio and Alexandrov-Kabadjov, 2004; Poesio et al., 2004a)

Although the results of the studies using the GNOME corpus mentioned above have been published in a number of papers, and although a detailed annotation manual was written and has been available on the Web for a few years (Poesio, 2000b), none of the previously published papers discusses in detail the goals of the annotation and the methodology that was followed, especially for the non-anaphoric aspects. In this paper we discuss the methods used to identify possible 'utterances,' the properties of NPs and discourse entities that were annotated, and (very briefly) anaphoric information.
.

## 2 The Data

Texts from three domains were (partially) annotated. The museum subcorpus consists of descriptions of museum objects and brief texts about the artists that produced them.[1] The pharmaceutical subcorpus is a selection of leaflets providing the patients with legally mandatory information about their medicine.[2] The GNOME corpus also includes tutorial dialogues from the Sherlock corpus collected at the University of Pittsburgh. Each subcorpus contains about 6,000 NPs, but not all types of annotation have been completed for all domains. All sentences, units and NPs have been identified, and all 'syntactic' properties of NPs (agreement feature and grammatical function). Anaphoric relations have been annotated in about half of the texts in each domain; and the more complex semantic properties (taxonomic properties, genericity, etc.) in about 25% of these texts. The total size of the annotated corpus is about 60K.

## 3 Identifying Utterances

In order to use a corpus to study salience, it is essential to find a way to annotate what in Center-

[1] The museum subcorpus extends the corpus collected to support the ILEX and SOLE projects at the University of Edinburgh (Oberlander et al., 1998).

[2] The leaflets in the pharmaceutical subcorpus are a subset of the collection of all patient leaflets in the UK which was digitized to support the ICONOCLAST project at the University of Brighton (Scott et al., 1998).

ing theory (Grosz et al., 1995) are called UTTERANCES, i.e., the units of text after which the local focus is updated. In most annotations concerned with salience, a predefined notion of utterance was adopted, typically sentences (Miltsakaki, 2002) or (finite) clauses (Kameyama, 1998). This approach, however, precludes using the corpus to compare possible definitions of utterance, one of the goals of the GNOME annotation (Poesio et al., 2004b).

In order to do this, we marked all spans of text that might be claimed to update the local focus, including sentences (defined as all units of text ending with a full stop, a question mark, or an exclamation point) as well as what we called (DISCOURSE) UNITS. Units include clauses (defined as sequences of text containing a verbal complex, all its obligatory arguments, and all postverbal adjuncts) as well as other sentence subconstituents that might be viewed as independently updating the local focus, such as parentheticals, preposed PPs, and (the second element of) coordinated VPs. Examples of clauses, verbal and non-verbal parentheticals, and preposed PPs marked as units follow; the parentheses indicate unit boundaries. (Sentence boundaries are not indicated.)

(1) a. **clausal unit with non-verbal parenthetical**: (It's made in the shape of a real object (– a violin))

b. **clausal unit with preposed PP and embedded relative clauses**: ((With the development of heraldry in the later Middle Ages in Europe as a means of identification), all (who were entitled (to bear arms)) wore signet-rings (engraved with their armorial bearings))

As example (1b) above illustrates, subordinate units such as clausal complements and relative clauses were enclosed within the superordinate unit. Subordinate units also include adjunct clauses headed by connectives such as *before, after, because* and clauses in subject position. In total, the texts used for the main study contain 505 sentences and more than 1,000 units, including 900 finite clauses.[3]

**Sentence and Unit Attributes** Sentences have one attribute, STYPE, specifying whether the sentence is declarative, interrogative, imperative, or exclamative. The attributes of units include:

- UTYPE: whether the unit is a main clause, a relative clause, appositive, a parenthetical, etc. The possible values for this attribute are `main`, `relative`, `such-as`, `appositive`, `parenthetical`, `paren-rel`, `paren-app`, `paren-main`, `subject`, `complement`, `adjunct`, `coord-vp`, `preposed-pp`, `listitem`, `cleft`, `title`, `disc-marker`.
- VERBED: whether the unit contains a verb.
- FINITE: for verbed units, whether the verb is finite or not.
- SUBJECT: for verbed units, whether they have a full subject, an empty subject (expletive, as in *there* sentences), or no subject (e.g., for infinitival clauses).

**Annotation Issues** Marking up sentences proved to be quite easy; marking up units, on the other hand, required extensive annotator training. The agreement on identifying the boundaries of units, using the $\kappa$ statistic discussed in (Carletta, 1996), was $\kappa = .9$ (for two annotators and 500 units); the agreement on features (2 annotators and at least 200 units) was as follows: UTYPE: $\kappa$=.76; VERBED: $\kappa$=.9; FINITE: $\kappa$=.81. The main problems when marking units were to identify complements, to distinguish clausal adjuncts from prepositional phrases, and how to mark up coordinated units. The main problem with complements was to distinguish non-finite complements of verbs such as *want* from the non-finite part of verbal complexes containing modal auxiliaries such as *get*, *let*, *make*, and *have*:

(2) a. (I would like (to be able to travel))

b. (I let him do his homework)

One problem that proved fairly difficult to handle (and which, in fact, we didn't entirely solve) was clausal coordination. The problem was to preserve enough structure to be able to compute the previous utterance, while preserving some basic intuitions about what constitutes a clause (roughly, that by and large clauses were text spans marked either by the presence of a semantically isolated verb or by punctuation / layout) which are essential for annotators and are needed to specify the values of attributes. This was relatively easy to do when two main clauses were coordinated; coordinated main clauses were marked as in (3a). However, it wasn't completely obvious what to do in the case of coordination within a subordinate clause, as in (3b). Because there weren't many such cases, rather than using the ⟨`unit`⟩ element with a special value for UTYPE as we did for coordinated NPs (which meant specifying all sorts of special values for attributes) we used a markup element called

[3] Our instructions for marking up such elements benefited from the discussion of clauses in (Quirk and Greenbaum, 1973) and Marcu's proposals for discourse units annotation (1999).

⟨unit-coordination⟩ to maintain the structure, and then marked up each clause separately, as shown in (3c) (the ⟨unit-coordination⟩ is marked with square brackets).

(3) a. (The Getty museum's microscope still works,) (and the case is fitted with a drawer filled with the necessary attachments).

b. (If you have any questions or are not sure about anything, ask your doctor or your pharmacist)

c. ((If [(you have any questions) or (you are not sure about anything)]), ask your doctor or your pharmacist)

The elements of text *not* marked up as units include: NPs, post-verbal and post-nominal PPs, non-verbal NP modifiers, coordinated VPs in case the second conjunct did not have arguments (4a), and quoted parts of text, when not reported speech (4b).

(4) a. (The oestradiol and norethisterone acetate are plant derived and synthetically produced)

b. (The inscription 'CHNETOC BASHLHKOC CPATHARHC')

**Layout** Our genres raised a few issues that, as far as we know, have not been previously discussed in the Centering literature. One such problem is what to do with layout elements such as titles and list elements, which can clearly serve as the first introduction of a CF and to move the CB. One example of title unit is unit (u1) in (5).

(5) (u1) Side effects

Side effects may occur when PRODUCT-Y is applied to large parts of the body,

We marked these layout elements as units, as in (6), but using the special value `title` of the attribute UTYPE (see above) so that we could test whether it was better to treat them as utterances or not.

(6) `<unit id="u1" utype="title">`Side effects`</unit>`

`<p><s stype="decl"><unit>` Side effects may occur `<unit>`when PRODUCT-Y is applied to large parts of the body, ... `</unit>` ... `</unit>` ... `</s>` ... `</p>`

**Problems with Attributes** The most difficult attribute to mark was UTYPE, and our main problem was to distinguish between relative clauses and parentheticals, since it's not always easy to tell whether a relative clause is restrictive or non-restrictive (see also (Cheng et al., 2001)). In the end, we adopted rules purely based on surface form (the presence or absence of a comma or other bracketing device). (See also (Quirk and Greenbaum, 1973).)

**Utterances and Propositions** The annotation of units has been shown useful to identify many of the atomic propositions expressed by a text, and was therefore used as a basis for studying text planning (Karamanis, 2003) and aggregation (Cheng, 2001).

## 4 Properties of Discourse Entities and their Realization

The main goal of the GNOME annotation was to study the factors that affect the realization of discourse entities, focusing on those entities realized as NPs. Hence, our main concern was to identify and to annotate the relevant properties both of discourse entities themselves and their realizations in a particular utterance (which we will call FORWARD LOOKING CENTERS, or CFs, following Centering's terminology). Both types of properties were annotated as properties of the ⟨ne⟩ element, used to mark up NPs in the corpus. Overall, we annotated 14 attributes of ⟨ne⟩ elements, specifying the syntactic and semantic properties of NPs and the semantic properties of the discourse entities they realize. We discuss these attributes in this section. We also annotated semantic relations between discourse entities, particularly when they express anaphoric relations. Anaphoric annotation is discussed in the next section.

### 4.1 Marking up NEs

The ⟨ne⟩ element is used to mark NPs, as in the following example (the attributes will be discussed below):

(7)

```
<unit finite='finite-yes' id='u3' utype='main'
      verbed='verbed-yes'>

<ne id="ne2" cat="poss-np" per="per3" num="sing"
      gen="neut" gf="subj" lftype="term"
      onto="concrete" ani="inanimate"
      deix="deix-no" count="undersp-count"
      structure="undersp-structure"
      generic="generic-no" loeb="sem-function">

  <ne id="ne3" cat="this-np" per="per3" num="sing"
      gen="neut" gf="gen" lftype="term"
      onto="concrete" ani="inanimate"
      deix="deix-yes" count="count-yes"
      structure="atom"
      generic="generic-no" loeb="pragm-function">
      This table's
  </ne>

</ne>
allow

<ne id="ne4" cat="bare-np" per="per3" num="plur"
gen="neut" gf="obj" lftype="term" onto="person"
ani="animate" deix="deix-no" count="count-yes"
structure="set" generic="generic-yes"  loeb="sort">
scholars </ne>

<unit finite='finite-no' id='u4' utype='complement'
verbed='verbed-yes'>
to link

<ne id="ne5" cat="pers-pro" per="per3" num="sing"
```

```
gen="neut" gf="obj" lftype="term" onto="concrete"
ani="inanimate" deix="deix-yes" count="count-yes"
structure="atom" generic="generic-no"
loeb="disc-function">  it  </ne>
...
```

The GNOME instructions for identifying NPs derive from those proposed in MATE (Poesio et al., 1999), in turn derived from DRAMA (Passonneau, 1997) and MUC-7 (Hirschman, 1998). An important difference between the instructions used for GNOME and those developed for MATE is that instead of attempting to get the annotators to recognize the NP that realize discourse entities and only mark those, in GNOME *all* NPs were marked with ⟨ne⟩ elements; the separate LF_TYPE attribute was used to distinguish between NPs with different types of denotations (see below). This change made the process of identifying nominal entities easier and potentially automatic (even though the identification of markables was still done by hand).

As in the case of units, the main problem with marking up NPs was coordination. Our approach was to use a separate ⟨ne⟩ element to mark up the coordinated NP, with type (CAT) value coord-np. We only used a coord-np element if two determiners were present, as in *((your doctor) and (your pharmacist))*. This approach was chosen because it limited the number of spurious coordinations introduced (in cases such as *this is an interesting and well-known example of early Byzantine jewellery*), but has the limitation that only one ⟨ne⟩ is marked in cases such as *Your doctor or pharmacist*.

### 4.2 Properties of all NPs

Some of the attributes of ⟨ne⟩ elements specify properties of all NPs, whether or not they realize a discourse entity. We discuss these first.

CAT The CAT attribute is used to mark NP type: whether the NP is a pronoun, a definite description, etc.. This attribute is only meant to provide a very surface-y classification, without attempting to group NPs in larger classes such as 'definite NP' and 'indefinite NP'. The one attempt to go beyond pure surface was the introduction of a distinction between definite descriptions that are really disguised proper names such as *the Beatles*, classified as CAT=the-pn, and all other definite descriptions, classified as the-np. The complete list of values for CAT is: a-np, another-np, q-np, num-np, meas-np, that-np, this-np, such-np, wh-np, poss-np, bare-np, pn, the-pn, the-np, pers-pro, poss-pro, refl-pro, rec-pro, q-pro, wh-pro, this-pro, that-pro, num-ana (for 'numerical anaphors' such as *one* in *I want one*), null-ana, gerund (for nominalized present participles such as *veneering furniture* in *the practice of veneering furniture*), coord-np, and free-rel (for 'free relatives' such as *what you need most* in *what you need most is a good rest*)).

The agreement on this attribute was pretty high, $\kappa = .9$; the one problem was the distinction between the-pn and the-np.

**Agreement features: NUM, PER, and GEN** These atributes are used to annotate features that are important to study pronoun interpretation: gender, number and person of NPs. Person and number were generally easy to annotate, but gender was very difficult because of the presence of many references to individual of unspecified gender, such as *the maker* in *the inventory gives neither the name of the maker nor the location*. This problem was solved by introducing a special undersp-gen value; indeed, underspecified values were provided for all attributes. The agreement values for these features were: GEN: $\kappa = .89$; NUM: $\kappa = .84$; PER: $\kappa = .9$.

GF This attribute was used to annotate the grammatical function of the NP, a property generally taken to play an important role in determining the salience of the discourse entity it realizes (Grosz et al., 1995). Our instructions for this attribute are derived from those used in the FRAMENET project ((Baker et al., 1998); see also http://www.icsi.berkeley.edu/~framenet/). The values are subj, obj, predicate (used for post-verbal objects in copular sentences, such as *This is (a production watch)*), there-obj (for post-verbal objects in *there*-sentences), comp (for indirect objects), adjunct (for the argument of PPs modifying VPs), gen (for NPs in determiner position in possessive NPs), np-compl, np-part, np-mod, adj-mod, and no-gf (for NPs occurring by themselves - eg., in titles). The agreement values for GF is $\kappa = .85$.

LF_TYPE Not all NPs realize discourse entities: some of them realize quantifiers (e.g., *each coffer* in *Each coffer has a lid*) or predicates (e.g., NPs in appositive position, such as *the oldest son of Louis XIV* in *The 1689 inventory of the Grand Dauphin, the oldest son of Louis XIV, lists a jewel coffer of similar form and decoration*. As said above, in the GNOME annotation all NPs are treated as markables, but the LF_TYPE attribute is used to indicate the type of semantic object denoted by an NP: term, quant or pred. Quantifiers were identified purely on the basis of the value of the CAT value: all NPs with CAT=q-np or q-pro should get a value of quant. A more complex test was used to identify

predicative NPs: three linguistic contexts in which NP are typically predicative were considered (appositions, postcopular position in *there*-sentences, and *become*-style sentences) but the annotators were explicitly asked to check whether the NP was used to express a property. Agreement was more tentative: $\kappa = .73$ (for two annotators, 200 NPs).

**Taxonomic information** Two semantic attributes capture information about the type of objects referred to (or quantifier over) by an NP. The first attribute, ONTO, was originally introduced to distinguish between gerunds (event nominalizations such as *letter-writing*) and bare plurals referring to concrete objects like *scholars*, both of which semantically denote collective objects (Link, 1983; Portner, 1992). Further distinctions were introduced to deal with 'difficult' objects, such as diseases; particular types of concrete objects such as medicines and persons were also singled out. Distinctions captured by the current set of values of ONTO include persons, medicines, other substances, other concrete objects; events, time intervals, or other abstract entities; spatial locations; and diseases. The agreement value for the latest version of ONTO was $\kappa = .8$ between two annotators, 200 NPs.

The second 'taxonomic' attribute, ANI, is used to annotate whether the objects referred to or quantifier over by an NP are animate or inanimate. This annotation was motivated by a number of studies suggesting that animacy plays an important role in salience (Prat-Sala and Branigan, 2000) and our own experiments suggesting that animacy is much more important than grammatical function, thematic roles, or order of mention in determining which entities are most likely to be pronominalized (Pearson et al., 2001). We also found that the discrepancy between the results of Gordon et al. (1999) and the findings of (Walker and Prince, 1996) can be explained in terms of animacy (Poesio and Nissim, 2001). Animacy was by far the easiest semantic attribute for our annotators: $\kappa = .92$.

### 4.3 Semantic properties of Discourse Entities

Semantic properties that may play a role in realization but only apply to discourse entities include:[4]

**Structure** Two attributes are used to indicate whether the discourse entity realized by an NP refers to a mass of certain substance or to countable objects (attribute COUNT) and, in case of countable objects, to an atom or a set (attribute STRUCTURE). These attributes were marked in order to study the factors leading to the realization of a discourse entity as a bare NP, in combination with the annotation of genericity discussed below: the reasoning being that it should only be possible to use bare singulars to realize a discourse entity described with mass nouns (as in *the ebeniste and his wife lived modestly in a five-room apartment ... with simple furniture*).[5]

The main reason for keeping the two attributes separate was that reaching agreement on STRUCTURE was fairly easy ($\kappa = .82$ at the second attempt) whereas COUNT was one of the most difficult attributes to mark–it took several iterations of changes to the instructions to achieve a $\kappa = .78$, and substantial revisions would probably still be useful. Nevertheless, given currently accepted views deriving from Link's work (1983), it would make more sense to merge the two attributes.

GENERIC This attribute is used to indicate whether the NP should be interpreted generically or not, which was thought to affect at least two types of discourse entity realizations: gerunds, that we took to be event types, and bare NPs, both singular and referring to substances (e.g., *ivory*) and plural. Annotating this information proved to be very difficult, which was not surprising because genericity is not yet a completely understood phenomenon. One complication is that there are two types of 'generic NPs': NPs referring to kinds, such as *The dodo* in *The dodo is extinct* (being extinct is not a property that can be predicated of individual dodos), and NPs used in generic statements, such as *Italians are good skiers* (a property of individual Italians) (Carlson and Pelletier, 1995). Although some NPs can only be used to express one or the other interpretation (e.g., * *A dodo is extinct*), many can be used in both ways (*Dodos are extinct*).

We started trying to make the very basic distinction between tokens and types one finds, e.g., in (Lyons, 1977), but even after numerous refinements we still encountered many problems. One of the problems our annotators had was whether to treat references to substances such as ivory and horn in examples like *This table's marquetry of ivory and horn* 'existentially,' i.e., as referring to the particular amounts of those substances used in the table, or 'generically', to refer to the kinds. In the end we decided to follow Carlson (1977) and to mark all of these examples as references to kinds, i.e., as generic. A second problem were quantifiers. Our annotators found it very hard to distinguish

[4] These attributes were only marked for about 25% of the corpus.

[5] Apart from the cases in which bare singulars are used to refer to substances, such as *the interiors of this pair of coffers are lined with tortoiseshell and brass*, the few discussed exceptions to this rule are expressions like *home* in *I went home*.

between quantified NPs used (non-generically) to quantify over a specific set of individuals at a particular spatio-temporal location, as in *Many lecturers went on strike (on March 16th, 2004)*, and quantifiers used in generic sentences, as in *Many lecturers went (habitually) on strike (during those years)*. The last version of the instructions (not yet added to the overall annotation manual) asked annotators to try to identify generic sentences before attempting to determine the value of the `GENERIC` attribute. With these instructions, we finally reached a reasonable agreement ($\kappa = .82$).

`LOEB` Poesio and Vieira (1998) found that of the 1,400 definite descriptions in their corpus, only about 50% were subsequent mention or bridging references, whereas 50% were first mentions. Of the first mentions, about half (i.e., 25% of the total) were what Hawkins (1978) would call 'larger situation' definites, i.e., definite descriptions like *the pope* whose referent is supposed to be part of shared knowledge; whereas the other half includes what Loebner (1987) calls SEMANTICALLY FUNCTIONAL definites, like *the first man on the Moon*. Loebner claimed that the paradigmatic case of definiteness are not anaphoric NPs, as suggested by familiarity theories such as Heim's (1982), but semantically functional ones such as *the first person ever to row across the Pacific on his own*. In order to test Loebner's theory and compare it with one based on familiarity, we annotated the NPs referring to discourse entities according to whether they were functional, relational, or sortal (Poesio, 2004a). We achieved good reliability on this attribute ($\kappa = .82$), and the results do suggest a much greater correlation between functionality and definiteness than between familiarity and definiteness (Poesio, 2004a).

## 5 Anaphora

The one aspect of the GNOME annotation that has been extensively discussed in previous papers is anaphoric annotation (Poesio, 2004b; Poesio et al., 2004b); we only discuss this aspect briefly here.

### 5.1 Annotating Discourse Models

Anaphoric annotation raises a number of difficult and, sometimes, unresolved semantic issues (Poesio, 2004b). As part of the MATE and GNOME projects, an extensive analysis of previously existing schemes for so-called 'coreference annotation,' such as the MUC-7 scheme, was carried out, highlighting a number of problems with such schemes, ranging from issues with the annotation methodology to semantic issues. Proposals for annotating 'coreference' such as (Hirschman, 1998) have been motivated by work on Information Extraction, hence the notion of 'coreference' used is very difficult to relate to traditional ideas about anaphora (van Deemter and Kibble, 2000). A distinctive feature of the GNOME annotation (and the MATE proposals from which they derive (Poesio, 2004b)) are explicitly based on the DISCOURSE MODEL assumption adopted almost universally by linguists (computational and not) working on anaphora resolution and generation (Webber, 1979; Heim, 1982; Kamp and Reyle, 1993; Gundel et al., 1993). This is the hypothesis that interpreting a discourse involves building a shared discourse model containing DISCOURSE ENTITIES that may or may not 'refer' to specific objects in the world, as well as the relations between these entities. The annotation for which the MATE scheme was developed–that we'll call here 'anaphoric annotation,' is meant as a partial representation of the discourse model evoked by a text.

### 5.2 Anaphoric Annotation in GNOME

For the GNOME corpus, we adopted a simplified version of the MATE scheme, as for our purposes it's not essential to mark all semantic relations between entities introduced by a text, but only those that may establish a 'link' between two utterances. So, for example, it was not necessary for us to mark a relation between the subject of a copular sentence and its predicate - e.g., between *the price of aluminum siding* and *$3.85* or *$4.02* in the example above.

In the GNOME corpus, anaphoric information is marked by means of a special ⟨`ante`⟩ element; the ⟨`ante`⟩ element itself specifies the index of the anaphoric expression (a ⟨`ne`⟩ element) and the type of semantic relation (e.g., identity), whereas one or more embedded ⟨`anchor`⟩ elements indicate possible antecedents.[6] (See (8).)

(8)
```
<unit finite='finite-yes' id='u227'>
 <ne id='ne546' gf='subj'>  The drawing of
   <ne id='ne547' gf='np-compl'>the corner cupboard
   </ne></ne>
 <unit finite='no-finite' id='u228'>,or more probably
   <ne id='ne548' gf='no-gf'> an engraving of
      <ne id='ne549' gf='np-compl'>it </ne></ne>
 </unit>,
...
</unit>
<ante current="ne549" rel="ident"> <anchor ID="ne547">
</ante>
```

Work such as (Sidner, 1979; Strube and Hahn, 1999), as well as our own preliminary analysis, suggested that indirect realization can play a crucial role in maintaining the CB. However, previous attempts at marking anaphoric information, particularly in the context of the MUC initiative, suggested that while agreement on identity relations is

[6] The presence of more than one ⟨`anchor`⟩ element indicates that the anaphoric expression is ambiguous.

fairly easy to achieve, marking bridging references is hard; this was confirmed by Poesio and Vieira (1998). For these reasons, and to reduce the annotators' work, we did not mark all relations. Besides identity (IDENT) we only marked up three associative relations (Hawkins, 1978): set membership (ELEMENT), subset (SUBSET), and 'generalized possession' (POSS), which includes part-of relations as well as ownership relations. We only marked relations between objects realized by noun phrases, excluding anaphoric references to actions, events or propositions implicitly introduced by clauses or sentences. We also gave strict instructions to our annotators limiting how much to mark.

As expected, we found a reasonable (if not perfect) 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 the 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. With 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 relations are marked by only one or the other annotator. So reaching agreement on this information involved several discussions between annotators and more than one pass over the corpus.

## 6 Automatically computing the Local Focus

The reader will have noticed that no attempt was done to directly mark up properties of the local focus - e.g., which discourse entity is the CB of a particular utterance. We found that it is much easier to annotate the 'building blocks' of a theory of the local focus, and then use scripts to automatically compute the CB. There are two advantages to this approach: first of all, agreement on the 'building blocks' is much easier to reach than agreement on the CB–in our preliminary experiments we didn't go beyond $\kappa = .6$ when trying to directly identify the CB using the definitions from (Brennan et al., 1987). And secondly, this approach makes it possible to compute the CB according to different ways of instantiating what we call the 'parameters of Centering' –e.g., ranking.

We developed such scripts for the work discussed in (Poesio et al., 2004b); they can be tested on the web site associated with that paper, `http://cswww.essex.ac.uk/staff/poesio/cbc/`. These scripts have been subsequently used to compute the CB in, e.g., (Poesio and Nissim, 2001; Poesio and Nygren-Modjeska, To appear).

## 7 Discussions and Conclusion

**Corpus consistency** The main lesson learned from this effort is that actually using a corpus is the best way both to ensure its correctness and to learn which types of information are most useful.

**Thematic Roles** One attribute on which we weren't able to reach acceptable agreement was the thematic role of an NP, which has been argued to be a better indicator of salience than grammatical function (Sidner, 1979; Stevenson et al., 1994); the agreement value in this case was $\kappa = .35$. Other groups however have shown that this can be done, e.g., in Framenet (Baker et al., 1998) and more recently in PropBank (Kingsbury and Palmer, 2002).

**Planned Revisions of the Scheme** A number of aspects of the annotation scheme used for the corpus could be improved. An obvious improvement would be to directly annotate predicates with their WordNet senses instead of annotating ONTO and animacy. We started doing this for the annotation of modifiers (Cheng et al., 2001), and developed an interface to WordNet, but too late to redo the whole corpus. Of the attributes, COUNT and GENERIC were the most difficult to annotate; further tests with these attributes could be useful.

**Automatic annotation** A substantial part of the annotation work required for GNOME now could (and should) be done automatically, or semi-automatically. This includes, most obviously, the identification of sentences and NPs, already done automatically in the VENEX corpus (Poesio, 2004b); and at least grammatical function, animacy, and countability could be automatically annotated in preliminary form with existing techniques, and then corrected by hand. We also plan to use the corpus to bootstrap techniques for automatic identification of uniqueness and gender.

## Acknowledgments

Special thanks to Janet Hitzeman, who collected the first subset of the museum domain for SOLE; to Renate Henschel, who completed the collection of the museum subset and wrote the first version of the annotation manual; to all our annotators; and to Mijail Alexandrov-Kabadjov and Nikiforos Karamanis, who identified a number of annotation problems. Most of this work was supported by the EPSRC project GNOME, GR/L51126/01.

## References


C. F. Baker, C. J. Fillmore, and J. B Lowe. 1998. The Berkeley FrameNet project. In *Proc. 36th ACL*.

S.E. Brennan, M.W. Friedman, and C.J. Pollard. 1987. A centering approach to pronouns. In *Proc. of the 25th ACL*.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comp. Linguistics*, 22(2):249–254.

G. N. Carlson and F. J. Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press.

G. N. Carlson. 1977. *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts, Amherst.

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

H. Cheng. 2001. *Modelling Aggregation Motivated Interactions in Descr. Text Generation*. Ph.D. thesis, Edinburgh.

P. C. Gordon, R. Hendrick, K. Ledoux, and C. L. Yang. 1999. Processing of reference and the structure of language: an analysis of complex noun phrases. *Language and Cognitive Processes*, 14(4):353–379.

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. K. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

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

I. Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, Univ. of Massachusetts at Amherst.

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

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

M. Kameyama. 1998. Intra-sentential centering. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, chapter 6, pages 89–112. Oxford.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

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

P. Kingsbury and M. Palmer. 2002. From Treebank to PropBank . In *Proc. of LREC*.

G. Link. 1983. The logical analysis of plurals and mass terms: A lattice- theoretical approach. In R. Bäuerle, C. Schwarze, and A. von Stechow, editors, *Meaning, Use and Interpretation of Language*, pages 302–323. Walter de Gruyter.

S. Loebner. 1987. Definites. *Journal of Semantics*, 4:279–326.

J. Lyons. 1977. *Semantics*. Cambridge.

D. Marcu. 1999. Instructions for manually annotating the discourse structures of texts. Unpublished manuscript.

E. Miltsakaki. 2002. Towards an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, 28(3):319–355.

J. Oberlander, M. O'Donnell, A. Knott, and C. Mellish. 1998. Conversation in the museum. *New Review of Hypermedia and Multimedia*, 4:11–32.

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

J. Pearson, R. Stevenson, and M. Poesio. 2000. Pronoun resolution in complex sentences. In *Proc. of AMLAP*, Leiden.

J. Pearson, R. Stevenson, and M Poesio. 2001. The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In M. Poesio, editor, *Proc. of SEMPRO-2001*. University of Edinburgh.

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

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 Inf. Structure, Disc. Structure and Disc. Semantics*.

M. Poesio and M. Nissim. 2001. Salience and possessive NPs: the effect of animacy and pronominalization. In *Proc. of AMLAP (Poster Session)*.

M. Poesio and N. Nygren-Modjeska. To appear. Focus, activation, and this-noun phrases. In A. Branco, R. McEnery, and R. Mitkov, editors, *Anaphora Processing*. John Benjamins.

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. In *Proc. of the ACL*.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. M. Hitzeman. 2004b. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3).

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

M. Poesio, 2000b. *The GNOME Annotation Manual, Fourth Edition*. Available from `http://www.hcrc.ed.ac.uk/~gnome`.

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

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

M. Poesio. 2004b. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston, May.

P. H. Portner. 1992. *Situation Theory and the Semantics of Propositional Expressions*. Ph.D. thesis, University of Massachusetts at Amherst.

M. Prat-Sala and H. Branigan. 2000. Discourse constraints on syntactic processing in language production. *Journal of Memory and Language*, 42(168–182).

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

D. Scott, R. Power, and R. Evans. 1998. Generation as a solution to its own problem. In *Proc. of the 9th INLG*.

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

R. J. Stevenson, R. A. Crawley, and D. Kleinman. 1994. Thematic roles, focus, and the representation of events. *Language and Cognitive Processes*, 9:519–548.

M. Strube and U. Hahn. 1999. Functional centering–grounding referential coherence in information structure. *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.

M. A. Walker and E. Prince. 1996. A bilateral approach to givenness. In J. Gundel and T. Fretheim, editors, *Reference Accessibility*, pages 291–306. John Benjamins.

B. L. Webber. 1979. *A Formal Approach to Discourse Anaphora*. Garland, New York.