# Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results

## **Massimo Poesio**

University of Edinburgh, HCRC and Informatics {Massimo.Poesio}@ed.ac.uk

# Abstract

We are annotating a corpus with information relevant to discourse entity realization, and especially the information needed to decide which type of NP to use. The corpus is being used to study correlations between NP type and certain semantic or discourse features, to evaluate hand-coded algorithms, and to train statistical models. We report on the development of our annotation scheme, the problems we have encountered, and the results obtained so far.

# 1. MOTIVATIONS

The goal of the GNOME project is to develop NP generation algorithms that can be used by real systems, with different architectures, and operating in realistic domains. As part of the project, we have been annotating a corpus with the syntactic, semantic and discourse information that is needed for different subtasks of NP realization, including the task of deciding on the most appropriate NP type to be used to realize a certain discourse entity (proper name, definite description, pronoun, etc.), and the task of organizing the additional information to be expressed with that discourse entity. We are using the annotated corpus to extract information useful to the development of hand-coded algorithms for the subtasks of NP realization we are focusing on, to develop statistical models of these subtasks, and to evaluate both types of algorithms. Conversely, we have been using the results of this evaluation to verify the completeness of our annotation scheme and to identify modifications. The annotation scheme used in our first corpus annotation exercise was discussed in (Poesio et al., 1999b); in this paper we present the modified annotation scheme that we developed as a result of that preliminary work, and discuss the problems we encountered when trying to annotate semantic and discourse information, and some issues that raise when annotating to evaluate generation algorithms as opposed to analysis ones.

The systems we are working with are the ILEX system developed at HCRC, University of Edinburgh (Oberlander et al., 1998),<sup>1</sup> and the ICONOCLAST system (Scott et al., 1998), developed at ITRI, University of Brighton. The ILEX system generates Web pages describing museum objects on the basis of the perceived status of its user's knowledge and of the objects she previously looked at; ICONOCLAST supports the creation of pharmaceutical leaflets by means of the WYSIWYM technique in which text generation and user input are interleaved.

## 2. DISCOURSE ENTITY REALIZATION

One of the problems encountered when trying to evaluate generation algorithms is that it's often difficult to isolate subtasks. Producing a sentence constituent syntactically classified as a noun phrase involves a number of tasks performed at different points of the generation process, and it's often not clear at which level certain decisions are best made. In order to be clear about the tasks we are studying it is necessary to make quite a few assumptions about generation and about the representations used at each level. In this section we briefly review our assumptions about NP generation in general and introduce the subtasks on which we focused in GNOME.

We assume an abstract architecture for NP generation roughly along the lines of that suggested by Dale, 1992) (see also (Cahill et al., 1999; Reiter and Dale, 2000)). According to Dale, the task of generation involves, first of all, the production of a tree-like TEXT PLAN encoding the information to be generated; the structure of the text plan encodes the rhetorical structure of the text, and its leaves are logical form representations of the propositions to be expressed. This text plan may subsequently be rearranged during an AGGREGATION phase. Next comes the REAL-IZATION phase, in which the logical forms associated with the leaves of the text plan are turned first into syntactic structures and then into strings. We further assume that the logical forms are Discourse Representation Structures (DRSs) (Kamp and Reyle, 1993; Heim, 1982); concretely, this means that we assume (i) that definites and indefinites are the expressions of terms rather than quantifiers, and (ii) that we assume that definites and indefinites are anaphorically accessible, but not all quantifiers necessarily are.

The task(s) we are concerned with in GNOME are part of the realization phase; i.e., the input to our algorithms includes pointers to elements of the logical form. Noun phrases appear in the generated text as the realization of at least three different types of logical form constituents:

- terms, which include referring expressions, as in *Jessie M. King* or *the hour pieces here*, but also non-referring terms such as *jewelry* or *different types of creative work*. Terms are called DISCOURSE ENTI-TIES in Discourse Representation Theory.
- **quantifiers**, as in *quite a lot of different types of creative work* or *nearly every day*
- **nominal predicates**, such as *an illustrator* in *She was an illustrator*.

<sup>&</sup>lt;sup>1</sup>The latest version of the system can be found at http://www.cstr.ed.ac.uk/cgi-bin/ilex.cgi.

Noun phrases can be **coordinated**, as in *The patches also contain oestradiol and norethisterone acetate* or *the inventory gives neither the name of the maker nor its original location*; we finesse the many issues raised by coordination by assuming a fourth type of logical form objects, **coordinations**.

We concentrate here on the task of realizing discourse entities, which we assume implemented by a procedure called realize-discourse-entity. realizediscourse-entity takes as input information that can be encoded in terms of feature structures as in (1). In particular, the value of the **sem** feature is a previously introduced discourse entity x, whose FILE CARD (Heim, 1982)a record containing the information about a discourse entity already introduced in discourse-contains the information that it is of type **woman** and it has name *Mary*, as in (2):

(1) 
$$\begin{bmatrix} syn : [cat : np] \\ sem : x \end{bmatrix}_{n_I}$$

(2) 
$$[x | woman(x), Mary(x)]$$

realize-discourse-entity also receives in input pointers to the text plan and to the feature structure representing all the other decisions already taken about the realization of the current node of the text plan; and outputs a modified feature structure in which some of the features encoding the abstract syntactic structure of the NP has been filled. As the syntactic structure of NPs is far from clear, we simply assume here that NPs have a flat structure with three list-valued features, det-sequence, premod, and postmod, and one slot head with value an object of type n.

$$syn : \left[ cat : np \right]$$

$$sem : \dots$$

$$det - sequence : \langle \dots \rangle$$

$$premod : \langle \dots \rangle$$

$$head : \left[ \right]_{n}$$

$$postmod : \langle \dots \rangle$$

In case realize-discourse-entity decides to realize the discourse entity using a proper name, its output can be represented as in (3), where we have assumed that proper names occupy the specifier position in noun phrases, as proposed in recent theories of noun phrase syntax such as (Abney, 1987)

(3) 
$$\left| \begin{array}{c} \operatorname{syn}: \left[ \operatorname{cat}: \operatorname{np} \right] \\ \operatorname{sem}: x \\ \operatorname{det} - \operatorname{sequence}: \left\langle \left[ \begin{array}{c} \operatorname{syn}: \left[ \operatorname{cat}: \operatorname{pn} \right] \\ \operatorname{sem}: \left[ |\operatorname{\mathbf{Mary}}(x) \right] \right] \right\rangle \right|_{\operatorname{np}} \\ \end{array} \right|_{np}$$

The task of realize-discourse-entity can be seen as consisting of two distinct subtasks. We use the term

NP TYPE DETERMINATION to indicate the process which decides whether to use, say, a pronoun, definite description, or proper name to refer to the discourse entity under consideration. Given the assumptions above about the syntactic representation of proper names and pronouns, this process amounts to a decision about the determiner to be used for a given NP, i.e., about the filler of the **det-sequence** slot (see (3)). The second task, which we will call NP MOD-IFIER DISTRIBUTION, involves deciding how to distribute the rest of the material specified in the semantic representation of the discourse entity among the remaining slots of the **np** feature structure. For example, given the file card:

$$[x | x$$
 is  $y$ , red $(x)$ , car $(x)$ ]

(where is is the equality statement in DRT and y is a previously introduced discourse entity) the information that discourse entity x is red would be part of the premodifier list, whereas the fact that x is a car would be specified by the head. Between them, these two processes specify how the four slots of the np type ought to be realized. In this paper, we focus on the work we did in GNOME on NP type determination.

# 3. DEVELOPING A SCHEME FOR NP REALIZATION

Our approach to developing an annotation scheme to study NP realization has been to come up with an initial set of features on the basis of linguistic theory and psychological results, and then to progressively refine this initial set of features on the basis of whether they could be annotated reliably and/or how useful they were, the usefulness being testified by the results of the preliminary algorithms we developed - in particular, of the statistical models of the task trained used the annotated corpus. In other words, we have been eliminating features which cannot be reliably annotated, or that don't seem to contribute much to the task, and adding new features when the results of the statistical model suggest that some distinction is missing. In this section, after a quick discussion of our corpus, we review the evidence that we used to define our first scheme, discuss our initial agreement study, and the results of our first attempt at building a statistical model of NP type determination.

#### The GNOME Corpus

The corpus we have collected for GNOME includes texts from both the domains we are studying. It contains texts in the museum domain, extending the corpus collected by the SOLE project (Hitzeman et al., 1998); and texts from the corpus of patient information leaflets collected for the ICONOCLAST project. The initial GNOME corpus (Poesio et al., 1999b) consisted of two subsets of about 1,500 NPs each; since then, the corpus has been extended and currently includes about 3,000 NPs in each domain. We are also adding texts from a third domain, tutorial dialogues.

The texts in the corpus contain examples of all types of NPs considered in our scheme, including quantified NPs, singular and plural bare-NPs with both generic and specific readings, nominalizations (*itching*, *reddening*) and complex modification (*This table's marquetry of ivory and horn*, *painted blue underneath*). BARE–NPs were the most common type of NP in the first GNOME corpus (700 in total, 22%), followed by definite descriptions (THE–NPs; 596; 18.7%), proper names (PN; 321; 10%), personal pronouns (PERS–PRO; 311; 9.8%) and NPs with the indefinite article a (A–NPs; 260; 8.19%).

## Semantic and Discourse Features That May Affect NP Type Determination

Even if in this first phase we focused on realizing discourse entities only, we still need to know for each NP in the corpus its semantic type (quantifier, discourse entity, predicate, or coordination). Coming to the features more strictly needed to determine the NP type chosen to realize discourse entities, and given the distribution of the corpus, we need first of all to worry about bare-NPs. One of the conditions under which (singular) bare NPs are used is when the object denoted is mass (cfr. \**a gold/a jewel* vs. *gold/\*jewel*); the other is when the NP is used to express a generic reference, as in *The cabinets de curiosites contained natural specimens such as <u>shells</u> and fossils.* 

Much work on NP generation has been devoted to studying the discourse factors that determine the type of NP chosen to realize a given discourse entity, and in particular whether it should be realized by a definite or an indefinite NP (Prince, 1992; Loebner, 1987; Gundel et al., 1993). Among the discourse properties of a discourse entity claimed to affect its form are

- Whether it is discourse new or old (Prince, 1992): e.g., a new jewel would be introduced by means of the indefinite *a jewel*, whereas for an already mentioned one the definite description *the jewel* would be used. This simple notion of familiarity was refined by Prince herself as well by Gundel *et al.* (Gundel et al., 1993).
- Whether it's hearer-new or hearer-old (Prince, 1992).
- Whether it is referring to an object in the visual situation or not: if so, a demonstrative NP may be used, as in *this jewel*.
- Whether it's currently highly salient or not, which may prompt the use of a pronoun. Properties that have been claimed to affect the salience of a discourse entity include: whether it's the current CENTER (CB) or not (Grosz et al., 1995), or more generally whether that entity is the TOPIC of the current discourse (Reinhart, 1981; Garrod and Sanford, 1983); its grammatical function; whether it's animated or not; its role; its proximity. (For a discussion of the effect of these and other factors on salience see (Poesio and Stevenson, To appear)).

According to Loebner (Loebner, 1987), the distinguishing property of definites is not familiarity (a discourse notion), but whether or not the predicate denoted by the head noun is functional (a semantic notion).

#### **Evaluating the Agreement on Feature Annotation**

Empirical studies of NP use typically involve a single annotator annotating her corpus according to her own subjective judgment (Prince, 1992; Passonneau, 1998). In order for the results of a corpus-based study to be replicable, it is, however, essential to show that more than one person understands the scheme (Passonneau and Litman, 1993; Carletta, 1996); this is particularly important with potentially subjective properties of discourse such as topic, and even more so if one is to propose the annotation scheme as something that other groups may use to train statistical models for their domain. The first step towards extracting an annotation scheme from the list of features just discussed involves therefore checking which of these can be annotated reliably.

Previous work (Poesio and Vieira, 1998) suggested that some of the information that has been suggested to influence NP type determination is hard to annotate reliably: e.g., whether an entity was hearer-new or hearer-old, or whether it was related to a previous entity by a relation other than identity. We thus came up with the following list of features of NPs that we felt may play a role in NP generation, and on whose reliability we didn't have previous negative results:

• The output feature, i.e., the type of NP (which we called cat). Our first specification of this attribute included the following values:

PERS-PRO POSS-PRO REFL-PRO Q-PRO WH-PRO THIS-PRO THAT-PRO ONE-ANA NULL-ANA PN POSS-NP THE-NP THIS-NP THAT-NP A-NP ANOTHER-NP BARE-NP Q-NP WH-NP NUM-NP MEAS-NP GERUND COORD-NP

These labels should all be self-explicatory. NUM-NP is used for numerical NPs such as *three years*; whereas MEAS-NP is used for NPs such as *4 mg of Product X*. GERUND is used for NPs such as *swimming* in *swimming should be avoided*.

- Syntactic attributes: num, per, gen (for GENder).
- Semantic attributes:
  - den to mark the semantic type of an NP i.e., whether it is a discourse entity, a quantifier, or a predicate, and whether it's generic or not;
  - Count, to mark whether an NP is countable or mass;
  - loeb, to mark up the functionality of the head predicate ( to verify Loebner's claims).
- Discourse attributes:
  - disc: whether the NP is discourse-new or discourse-old;
  - deix: whether the NP is deictical or not;
  - cb: whether the NP is the current CB (Grosz et al., 1995);
  - top: whether the NP is the current topic or not .
- The following attributes were only marked if the NP was the antecedent of another NP (see below):

- ani: whether the NP is animate or inanimate;
- fun: grammatical FUNction of the NP;
- role: its thematic ROLE;
- the clause type (ct and mood (mood) of the clause in which it occurs.

We also marked anaphoric relations, and annotated the following properties of these relations:

- The proximity of the antecedent to the anaphor (prox);
- The relation between the antecedent and the anaphor (rel), similar to the ref attribute in the MUC scheme (Chinchor, 1997), with values IDENT, SUBSET, etc.

We ran a partial reliability study of these features by having two of the developers of the scheme independently annotate a subset of our corpus including 700 NPs and computing their agreement by means of the K statistic discussed by Carletta, 1996). A value of K between .8 and 1 indicates good agreement; a value between .6 and .8 indicates some agreement. The K results for the NP features were as follows:

Attribute	K Value
cat	.92
ani	.88
disc	.72
fun	.68
loeb	.63
cb	.6
ct	.51
den	.456
role	.42
top	.375

Concerning the agreement on anaphoric relations, we observed a fairly good agreement on identifying the antecedents of direct anaphoric references, but worse agreement on bridging references, as expected on the basis of the results in (Poesio and Vieira, 1998). The agreement on the features of anaphoric relations features in the cases when both annotators marked them, was as follows:

Attribute	K Value
prox	.61
rel	.6

Of the features of noun phrases, ct, den, role and top were the ones that most clearly failed to reach the level of minimum agreement. The disagreements among our annotators had three main causes. In some cases the problem was that certain notions are intrinsically difficult to define: this is the case, for example, of trying to annotate for topic, or for the thematic role of an entity. In order to be able to annotate this informatin, substantial theoretical work would be needed. In other cases, and in particular with the semantic attribute den, the problem was that different semantic analyses of certain NPs have been proposed in the literature–in (4a), for example, *long periods* can be analyzed either as a quantifier or as a kind–so that two annotators assuming two different theories could disagree. This forced us to abandon our attempt of not committing to any specific semantic theory, and to commit to DRT instead (where, e.g., bare plurals are systematically analyzed as terms rather than quantifiers). Finally, some NPs were ambiguous–e.g., in (4b), it's not clear whether *tortoise shell and brass or pewter* refers to these materials in the abstract or to the specific tokens used in the object being discussed–and the semantic type of other NPs was difficult to characterize, as in (4c).

- (4) a. Infants and children must not be treated continuously with Nerisone for long periods.
  - b. *The interiors of this coffer are lined with* <u>tortoise shell and brass</u> or pewter.
  - c. ... each decorated using a technique known as premiere partie marquetry, a pattern of brass and pewter on a tortoiseshell ground ...

As a consequence of these problems, we decided to eliminate from the scheme used for the preliminary annotation used to train our first statistical models (below) the least reliable features (top and role), and to drastically revise the manual, especially the instructions for annotating the two attributes with next lowest reliability, (den and loeb). This left us with the syntactic features cat, num, per, and gen; the semantic features count, den, and loeb; and the discourse features deix, disc, and cb.

#### Evaluating the Scheme by Building a Statistical Model

After this preliminary test of reliability, we proceeded to evaluate the complenetess of the remaining set of NP features by annotating the NPs in our corpus (not the anaphoric relations) using this smaller set of attributes, and using this annotation to build a statistical model of the process of NP type determination - i.e., the process by which the value of cat is chosen on the basis of the values of the other features. We tried both the Maximum Entropy model (Berger et al., 1996) as implemented by Mikheev (Mikheev, 1998) and the CART model of decision tree construction (Breiman et al., 1984); the results below were obtained using CART. The model was evaluated by comparing the label it predicted on the basis of the features of a given NP with the actual value of cat for that NP, performing a 10-fold cross-validation. The model achieved a 70% accuracy, against a baseline of 22% (if the most common category, BARE-NP, were chosen every time.) The accuracy for the main types of NPs was as follows:

CAT	Accuracy
PN	97.5%
PERS-PRO	94.1%
Q-NP	89.43%
BARE-NP	82.7%
THE-NP	81.10%
A-NP	68%
POSS-NP	46%

Of the remaining classes of NPs, our model gets THIS-NPs, POSS-PROs and GERUNDs mostly wrong; for the

Sys cl.:	PersPre	D	PossPr	0	Th	eNP	1	ThisNP
PersPro	28		0			0		0
PossPro	15		0			0		0
The-NP	0		0			50		0
ThisNP	0		0			7		1
A-NP	0		0			0		0
BareNP	0		0			7		0
Gerund	0		0			0		0
Sys cl.:	ANP	F	BareNP	G	ler	Tota	l	Perc
PersPro	0		0		0	28		100%
PossPro	0		0		0	15		0%
The-NP	0		0		0	60		83.3%
ThisNP	0		0		0	10		10%
A-NP	20		2		0	22		90.9%
BareNP	4		54		0	68		79.4%
Gerund	0		5		1	6		16.7%

Table 1: Comparison between the class of an NP as specified by the annotation (vertical dimension) and the class assigned to it by the system (horizontal).

other classes there aren't enough data to get significant results. Table 1 illustrates the most interesting classification errors on one of the test sets of the cross validation.

In the case of POSS-NPs, what is missing from the annotation is simply the information that the object denoted by the NP is 'owned' by some other entity; this information will become available once anaphoric relations have been annotated, since this annotation would also include information that there is a possession relation between the NP as a whole and the possessor.

The problem with GERUNDs is that they tend to be classified by the system as bare-NPs; this is because both types of NPs tend to denote types rather than tokens - types of events in the case of gerunds, types of 'concrete' individuals in the case of other bare NPs - but the current annotation scheme does not specify whether an entity denotes a set of events or a set of concrete individuals. What is missing from the scheme in this case is an indication of whether the object denoted by the discourse entity is an event.

The most complex problem to fix is that of THIS-NPs: here the reason for the misclassification is that THIS-NPs are used in our texts not only to refer to pictures or parts of them, but also to refer to abstract objects introduced by the text, as in the following examples:

- (5) a. A great refinement among armorial signets was to reproduce not only the coat-of-arms but the correct tinctures; they were repeated in colour on the reverse side and the crystal would then be set in the gold bezel. Although the engraved surface could be used for impressions, the colours would not wear away. The signet-ring of Mary, Queen of Scots (beheaded in 1587) is probably the most interesting example of this type;
  - b. The upright secrétaire began to be a fashionable form around the mid-1700s, when letter-writing became a popular past-time. The marchands-merciers were quick to respond to this demand,

Again, we expect to be able to improve the results for this class once we have completed the annotation of antecedent relations; however, it's not clear that this type of references can be annotated reliably.

### 4. THE NEW SCHEME

We are currently in the process of reannotating the corpus from scratch according to a new annotation scheme developed on the basis of the tests just discussed (reliability and/or incompleteness of information). We discuss the new scheme in this section. In addition, we are adding information about NP modification; this aspect of the new annotation won't be discussed here. For reasons of space, only a brief discussion is possible in particular, we won't be able to discuss in detail the instructions given to annotators; the complete instructions are available at http://www.hcrc.ed.ac.uk/ ~gnome/anno\_manual.html.

#### Annotation Methodology

One of the important lessons taught by our first attempts at annotation was that doing all of this annotation at the same time was too much, and that it was essential to order the annotation process so that, for example, information about clauses would be available when annotating for grammatical function. So we split the task as follows:

- 1. First mark up the layout
- 2. Then identify and annotate the units (see below)
- 3. Then identify NPs and mark their syntactic attributes
- 4. Then annotate anaphoric information
- 5. Then annotate the more complex semantic and discourse information.

#### Markup Language

Our annotation scheme is XML-based, which allows us to make use of the suite of XML manipulation tools developed by the Language Technology Group of the University of Edinburgh (http://www.ltg.ed.ac.uk/software/).

The basis for our annotation are a rather minimal set of layout tags, identifying the main divisions of texts, their titles, figures, paragraphs, and lists. Also, as a result of the reliability studies discussed below and of our first annotation effort, we decided to also mark up units of text that may correspond to rhetorical units in our second annotation, using the tag  $\langle unit \rangle$ .

An important feature of the scheme is that the information about NPs is split among two XML elements, as in the MATE scheme for coreference (Poesio et al., 1999a). Each NP in the text is tagged with an  $\langle ne \rangle$  tag, as follows:

(6) <ne ID="ne07" ... >
Scottish-born, Canadian based jeweller,
Alison Bailey-Smith</ne>
...
<ne ID="ne08"> <ne ID="ne09">Her</ne>
materials</ne>

the instructions for identifying the  $\langle ne \rangle$  markables are derived from those proposed in the MATE project scheme for annotating anaphoric relations (Poesio et al., 1999a), which in turn were derived from those proposed by Passonneau

(Passonneau, 1997) and in MUC-7 (Chinchor and Sundheim, 1995).

Anaphoric relations are annotated by means of a separate  $\langle \texttt{ante} \rangle$  element specifying relations between  $\langle \texttt{ne} \rangle$ s, also as proposed in MATE. An  $\langle \texttt{ante} \rangle$  element includes one or more  $\langle \texttt{anchor} \rangle$  element, one for each plausible antecedent of the current discourse entity (in this way, ambiguous cases can be marked). E.g., the anaphoric relation in (6) between the possessive pronoun with ID = "ne09" and the proper name with ID = "ne07" is marked as follows:

The information relevant to the task of modifier order determination is marked up by associating a (mod) tag with each NP modifier inside a (ne) tag:

## (Discourse) Units

The problems encountered with the cb, fun, and prox features, especially, suggested the need for marking up sentences and potential rhetorical units / centering theory utterances before marking up certain types of information about NPs such as grammatical function. The instructions for marking up units were in part derived from (Marcu, 1999); for each  $\langle unit \rangle$ , the following attributes were marked:

- **utype**: whether the unit is a main clause, a relative clause, appositive, a parenthetical, etc.
- verbed: whether the unit contains a verb or not.
- 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).

The agreement on identifying the boundaries of units was K = .9; the agreement on features was follows:

Attribute	K Value
utype	.76
verbed	.9
finite	.81
subject	.86

This part of the annotation has now been completed. The main difficulties we observed had to do with assigning an utterance type to parenthetical sentences.

### NEs

A few attributes were dropped from the scheme for marking up noun phrases used in the previous annotation effort: among these in particular disc (redundant once antecedent information was marked up) and cb (which could be automatically derived from the information about grammatical function and units). All attribute definitions and instructions were revised, in particular, the **den** attribute was drastically revised by separating off information about the logical form type of an NP (quantifier, term, etc) from the information about genericity. New attributes were introduced to annotate information about the abstractness or concreteness of an object, and about its semantic plurality or atomicity. The revised list of information annotated for each NP includes:

- The output feature, cat (slightly revised and with better instructions)
- The other 'basic' syntactic features, num, per, and gen (as in the previous scheme)
- A feature gf specifying its grammatical function;
- The following semantic attributes:
  - ani: whether the object denoted is animate or inanimate
  - count: whether the object denoted is mass or count
  - lftype: one of quant, term, pred, coord
  - generic: whether the object denoted is a generic or specific reference
  - Onto: whether the object denoted is concrete, an event, a temporal reference, or another abstract object
  - structure: whether the object denoted is atomic or not
- The following discourse attributes:
  - deix: whether the object is a deictic reference or not
  - unique: whether the description used allows the reader to characterize the object as unique

A number of NP properties (e.g., familiarity) can be derived from the annotation of anaphoric information (below); in addition, a few properties of NPs are automatically derived from other sources of information - e.g., the type of layout element in which the NP occurs (in titles, bare-nps are often used) and whether a particular NP has uniquely distinguishing syntactic features in a given unit. We haven't yet completed all the agreement studies for NP features; the results that we do have are as follows:

Attribute	K Value
ani	.81
cat	.9
deix	.81
gen	.89
gf	.85
lftype	.74
num	.84
per	.9

(One interesting point to note here is that agreement on lftype is actually quite high (90%), but because TERMs are so prevalent, chance agreement is also very high.)

#### **Antecedent Information**

Previous work, particularly in the context of the MUC initiative, suggested that while it's fairly easy to achieve agreement on identity relations, marking up bridging references is quite hard; this was confirmed, e.g., by (Poesio and Vieira, 1998). The only way to achieve a reasonable agreement on this type of annotation, and to contain somehow the annotators' work, is to limit the types of relations annotators are supposed to mark up, and specify priorities. We are currently experimenting with marking up only four types of relations, a subset of those proposed in the 'extended relations' version of the MATE scheme (Poesio et al., 1999a) (which, in turn, derived from Passonneau's DRAMA scheme (Passonneau, 1997): identity (IDENT), set membership (ELEMENT), subset (SUBSET), and 'generalized possession', including part-of relations.

In addition, given our interests we had to be quite strict about the choice of antecedent: whereas in MUC it is perfectly acceptable to mark an 'antecedent' which *follows* a given anaphoric expression, in order, e.g., to compute the CB of an utterance it is necessary to identify the *closest previous* antecedent.

As expected, we are achieving 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. On the other hand, only 22% of bridging references were marked in the same way by both annotators; although our current scheme does limit the disagreements on antecedents and relations (only 4.8% relations are actually marked differently) we still find that 73.17% of relations are marked by only one or the other annotator.

#### 5. DISCUSSION

There are some pretty obvious omissions in the work done so far. Even if we only consider the task of NP type determination, there are a number of features whose impact we haven't been able to study so far, in some cases because they proved very hard to annotate. We already discussed two such examples, topichood and thematic roles; another potentially important source of information about the decision to pronominalize, rhetorical structure, is even harder to annotate. We would like to be able to annotate some types of scoping relations as well, especially the cases in which an NP is in the scope of negation as this may license the use of polarity-sensitive items such as any. Another important factor is the role of the information which the text planner has decided to realize: e.g., once the text planner has decided to generate both the proper name of discourse entity x, Alphonse Mucha, and the fact that x is a Czech painter, the decision to use the THE-NP the Czech painter Alphonse Mucha is more or less forced on us. And of course, nothing in the scheme discussed above allows us to study the conditions under which a generator may decide to produce a quantifier or a coordinated NPs.

Among the issues raised by this work, an important one is how much of the information that we annotated by hand could be automatically extracted. We believe that a lot of the syntactic information we rely on ( $\langle \text{unit} \rangle$  and  $\langle \text{ne} \rangle$ identification,  $\langle \text{unit} \rangle$  attributes, basic syntactic attributes of  $\langle \text{ne} \rangle$ ) could be extracted automatically using recent advances in robust parsing; this would already cut down the amount of work considerably. The problem is what to do with semantic information: e.g., whether suitable approximations could be found.

Another important question is whether our characterization of NP realization is plausible. One could imagine two quite distinct objections. On the one hand, some could argue that we conflated into NP type determination two very different tasks: determining whether to use a numeral or not, and whether to use a definite article or not. Conversely, it could be argued that NP type determination goes hand-in-hand with content determination, as we saw above, and the two problems can only be attacked simultaneously. We are trying to address the general architectural problem by making sure that our algorithms can be embedded in real systems, being developed independently by other groups. The problem with the second type of objections is that it's very difficult to study content determination. This is because of a more general problem with the methodology we are using: there is a mismatch between what a system knows and what an annotator may know about an objecti.e., between the features that a generation system may use and the features that can be annotated, and it's not clear this mismatch can be resolved.

For one thing, the need to choose features that can be annotated reliably imposes serious constraints: features that a generation system can easily set up by itself (e.g., the ILEX system keeps track of what it thinks the current topic is) can be difficult for two annotators to annotate in the same way. Second, some information that a generation system can use when deciding on the type of NP to generate may simply be impossible to annotate. For example, we already seen that the form of an NP often depends on how much information the system intends to communicate to the user about a given entity, or how much information the system believes the user has. In order to build a model of this decision process, we would need to specify for each NP how much information it conveys, and of what type; it's not at all clear that it will be feasible to do this by hand, except in domains in which the annotator knows everything that there is to know about a given object (see, e.g., Jordan's work on the COCONUT domain (Jordan, 1999)).

Conversely, some information that can be annotated indeed, that is easy to annotate - may not be available to some systems. E.g., we do not know of any system with a lexicon rich enough to specify whether a given entry is functional or not. A solution in this case may be to develop algorithms to extract this information from an annotated corpus, or perhaps just using the syntactic distribution of the predicate as an indication (e.g., a predicate X occurring in a *the X of Y* construction may be functional).

In other words, we believe that the present work is only a first step towards developing an appropriate methodology for empirical investigation and evaluation of generation algorithms, which we nevertheless feel will become more and more necessary.

#### Acknowledgments

I wish to thank the other members of the GNOME project, without whom this work would not have been possible– in particular Hua Cheng, Kees van Deemter, Renate Henschel, Rodger Kibble, and Rosemary Stevenson. The corpus is being annotated by Debbie De Jongh, Ben Donaldson, Marisa Flecha-Garcia, Camilla Fraser, Michael Green, Shane Montague, Carol Rennie, and Claire Thomson. I also wish to thank Janet Hitzeman, Alistair Knott, Chris Mellish, Johanna Moore, and Jon Oberlander. The GNOME project is supported by the UK Research Council EPSRC, GR/L51126. Massimo Poesio is supported by an EPSRC Advanced Research Fellowship.

## 6. References

- Abney, S. P., 1987. *The English Noun Phrase in its Sentential Aspect*. Ph.D. thesis, MIT, Cambridge MA.
- Berger, A., S. Della Pietra, and V. Della Pietra, 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72.
- Breiman, L., J. H. Friedman, R. A Olshen, and C. J. Stone, 1984. *Classification and Regression Trees*. Chapman and Hall.
- Cahill, L., C. Doran, R. Evans, C. Mellish, D. Paiva, M. Reape, D. Scott, and N. Tipper, 1999. In search of a reference architecture for NLG systems. In *Proc. European Workshop on Natural Language Generation*.
- Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Chinchor, N. A., 1997. Overview of MUC-7/MET-2. In Proc. of the Seventh Message Understanding Conference (MUC-7). Available at http://www.muc.saic.com/proceedings/ muc\_7\_proceedings/overview.html.
- Chinchor, N. A. and B. Sundheim, 1995. Message Understanding Conference (MUC) tests of discourse processing. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.*
- Dale, R., 1992. Generating Referring Expressions. Cambridge, MA: The MIT Press.
- Garrod, S. C. and A. J. Sanford, 1983. Topic dependent effects in language processing. In G. B. Flores D'Arcais and R. Jarvella (eds.), *The Process of Language Comprehension*. Chichester: Wiley, pages 271–295.
- Grosz, B. J., A. K. Joshi, and S. Weinstein, 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225.
- Gundel, J. K., N. Hedberg, and R. Zacharski, 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.
- Heim, I., 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.
- Hitzeman, J., A. Black, P. Taylor, C. Mellish, and J. Oberlander, 1998. On the use of automatically generated discourse-level information in a concept-to-speech synthesis system. In Proc. of the International Conference on Spoken Language Processing (ICSLP98), Australia.

- Jordan, P., 1999. An empirical study of the communicative goals impacting nominal expressions. In R. Kibble and K. van Deemter (eds.), *Proc. of the ESSLLI workshop on The Generation of Nominal Expressions*. Utrecht: University of Utrecht, OTS.
- Kamp, H. and U. Reyle, 1993. From Discourse to Logic. Dordrecht: D. Reidel.
- Loebner, S., 1987. Definites. *Journal of Semantics*, 4:279–326.
- Marcu, D., 1999. Instructions for manually annotating the discourse structures of texts. Unpublished manuscript, USC/ISI.
- Mikheev, A., 1998. Feature lattices for maximum entropy modeling. In *Proc. of ACL-COLING*. Montreal, CA.
- Oberlander, J., M. O'Donnell, A. Knott, and C. Mellish, 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–32.
- Passonneau, R., 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.
- Passonneau, R., 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In M. A. Walker, A. K. Joshi, and E. F. Prince (eds.), *Centering Theory in Discourse*, chapter 17. Oxford University Press, pages 327–358.
- Passonneau, R. and D. Litman, 1993. Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL*.
- Poesio, M., F. Bruneseaux, and L. Romary, 1999a. The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker (ed.), *Proc. of the ACL Workshop on Standards and Tools for Discourse Tag*ging.
- Poesio, M., R. Henschel, J. Hitzeman, R. Kibble, S. Montague, and K. van Deemter, 1999b. Towards an annotation scheme for Noun Phrase generation. In B. Krenn H. Uszkoreit, T. Brants (ed.), *Proc. of the EACL workshop on Linguistically Interpreted Corpora (LINC-99).*
- Poesio, M. and R. Stevenson, To appear. Salience: Computational Models and Psychological Evidence. Cambridge and New York: Cambridge University Press.
- Poesio, M. and R. Vieira, 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Prince, E. F., 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann (eds.), *Discourse description: diverse analyses of a fundraising text*. John Benjamins, pages 295–325.
- Reinhart, T., 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1).
- Reiter, E. and R. Dale, 2000. *Building Natural Language Generation Systems*. Cambridge University Press.
- Scott, D., R. Power, and R. Evans, 1998. Generation as a solution to its own problem. In Proc. of the 9th International Workshop on Natural Language Generation. Niagara-on-the-Lake, CA.