Towards An Annotation Scheme For Noun Phrase Generation

M. Poesio, R. Henschel, J. Hitzeman,[†] R. Kibble,[§] S. Montague, and K. van Deemter[§]

HCRC, **University of Edinburgh, Edinburgh, Scotland**
{poesio,henschel,shane}@cogsci.ed.ac.uk
[†]**Centre for Speech Technology Research and** HCRC, **University of Edinburgh**
J.Hitzeman@ed.ac.uk
[§]ITRI, **University of Brighton, Brighton, UK**
{Rodger.Kibble,Kees.Van.Deemter}@itri.bton.ac.uk

Abstract

We are studying the feasibility of annotating a corpus with information relevant to NP generation - specifically, the information needed to decide which type of NP to use. Such a corpus might be used just to study correlations between NP type and certain semantic or discourse features, or 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

As part of the GNOME project, whose goal is to develop general-purpose NP generation algorithms, we are experimenting with annotating our corpus with the syntactic, semantic and discourse information that is needed to decide on the type of NP to be used to realize a given discourse entity–e.g., whether a proper name should be used, a definite description, or a pronoun: the effect of this variation being illustrated by examples such as (1).

(1) a. *Jessie M. King was a Liberty artist. Jessie M. King was not just a jewelry designer. Jessie M. King was an illustrator. Jessie M. King did quite a lot of different types of creative work. Jewelry is just part of it. The four pieces here show four quite distinct aspects of Jessie M. King's creative work.*

b. *Jessie M. King, the Liberty artist, was not just a jewelry designer. She was an illustrator, and did quite a lot of different types of creative work; jewelry is just part of it. The four pieces here show four quite distinct aspects of King's creative work.*

We expect this annotation to be useful at least to get some sort of evaluation of the completeness of the set of features used by our NP generation algorithms; we are also experimenting with developing statistical models of aspects of the NP generation task.

In this paper, we describe our experience so far with this kind of annotation. We are currently concerned with the process by which the type of a NP is determined, on the basis of semantic and discourse information about the discourse entity realized by that NP, and not with the task of deciding what type of information the NP should include, often discussed in the NP generation literature (Dale and Reiter, 1995).

2 ANNOTATING FOR ANALYSIS VERSUS ANNOTATING FOR GENERATION

The annotation schemes developed for evaluating coreference resolution algorithms, such as the scheme developed for MUC-7 (Chinchor and Sundheim, 1995), the DRAMA scheme proposed by (Passonneau, 1997), and the MATE project scheme for annotating anaphoric relations (Davies et al., 1998; Poesio et al., 1999) - provide some of the information needed to decide on the type of NP to use to realize a discourse entity. These schemes specify which anaphoric relations to annotate in a text, and how: in the SGML-based MUC-7 scheme, for example, each NP is annotated with an index and, possibly, a `REF` attribute specifying whether that NP has an antecedent in discourse. The scheme also includes instructions to identify the NPs that ought to be annotated (MARKABLES), excluding, e.g., expletives, or proper names occurring in premodifier position (as in *the Getty museum*).

One could use the `REF` attribute of the MUC scheme to draw a rough distinction between first-mention and subsequent-mention entities, using indefinite expressions for the former and definites for the latter; the scheme does not, however, allow us to make finer-grained distinctions between NPs - e.g., between bare singulars and a-nps, or between definite descriptions, proper names, and pronouns.[1]

3 OUR ANNOTATION SCHEME

Semantic and Discourse Features That May Affect NP Form

We are assuming an architecture for NP generation roughly along the lines of that suggested by (Dale, 1992), where the three main sources of information accessed during NP generation are a knowledge base in the form of a semantic network, a discourse model, and information about the decisions already taken by the generation system as a whole concerning the sentence of which the NP to be generated is meant to be a constituent.[2]

Much work on NP generation has been devoted to studying the discourse factors that determine the form of an NP, and in particular whether it should be realized by a definite or an indefinite NP: among the discourse properties of a discourse entity claimed to affect its form are

- Whether it is discourse new or old (Prince, 1981): 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.

- 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 suggest 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 (Garrod and Sanford, 1983); its grammatical function; whether it's animated or not; its role; its closeness (the claims that these and other factors contribute to salience are reviewed in (Poesio and Stevenson, To appear)).

Among the semantic properties of a discourse entity that may affect the way it is going to be realized as an NP, we have:

- Its semantic type: quantified objects are generally realized as quantifiers,[3] 'types' or 'kinds' are often referred to by means of bare plurals such as *jewels in this style*, and tokens or individuals are referred to by proper names, definite NPs, or indefinite NPs, depending on other semantic or discourse properties.

- Whether the object it denotes is mass or count: this is going to affect, among other things, whether a bare NP is going to be chosen–cfr. **a gold/a jewel, gold/*jewel.*

- According to (Loebner, 1987), the distinguishing property of definites is not familiarity (a discourse notion), but whether or not the object described denotes a function (a semantic notion).

The Annotation Scheme: General Features

Our annotation scheme is SGML-based, like the one used in MUC. The basis for our annotation are texts already annotated with a rather minimal set of layout tags, identifying the main divisions of texts, titles, figures, paragraphs, and lists.

An important feature of the scheme is that the information about NPs is split among two SGML elements. Each NP in the text is tagged with an NE tag, as follows:

(2)
```
<NE ID="07" ...
(other attributes, discussed below)>
Scottish-born, Canadian based
jeweller, Alison Bailey-Smith</NE>
...
<NE ID="08">
<NE ID="09">Her</NE>
materials</NE>
```

(the instructions for identifying the markables are derived from those proposed in the MATE project scheme for annotating anaphoric relations (Davies et al., 1998), which in turn were derived from those proposed by (Passonneau, 1997) and in MUC-7 (Chinchor and Sundheim, 1995)).

[1]In fact, the scheme would have to be slightly modified even to make the distinction between first mention and subsequent mention properly. In its current form, the scheme is order-independent: when an object is mentioned twice in a text, the annotator is allowed to tag the first mention with a REF attribute, with value the index of the second mention (van Deemter and Kibble, 1999).

[2]Details about the existing interface between the text planner and the NP planner in ILEX can be found in (O'Donnell et al., 1998).

[3]We are assuming a semantic network in which certain types of quantified statements can be expressed; we will not discuss here whether semantic networks are the appropriate way to represent quantification.

In addition, anaphoric relations are also annotated by means of a separate ANTE element specifying relations between NEs, also as proposed in MATE.[4] E.g., the anaphoric relation in (2) between the possessive pronoun with ID 09 and the proper name with ID 07 is marked as follows:

(3)
```
<ANTE CURRENT="09" SOURCE="07"
REL="ident"
 .... (other attributes)>
```

Some of the attributes relevant for NP generation are associated with the NE elements; other attributes –the NP properties that are only relevant when an NP serves as an antecedent–are associated with ANTE elements.

Initial Set of Features

A useful property of most statistical algorithms for creating language models is that one can encode every 'input' feature one thinks may have an impact on a given decision; it is then up to the system to find the best correlation between input and output features. The only factors that limit the annotation effort are, therefore, time and whether a given input feature can be encoded reliably. Our approach in choosing a scheme was, therefore, to begin with a fairly extensive set of attributes, narrowing it down to those attributes that could be reliably annotated.

On the basis of the considerations discussed above, we came up initially with the following list of features of NPs that we felt may play a role in NP generation:

- NE features:
  - The output feature, CAT, with 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 NUMWD-NP   NUMFIG-NP
    GERUND COORD-NP
    ```
  - Syntactic attributes: NUM, PER, GEN
  - Semantic attributes: DEN, LOEB
  - Discourse attributes: CB, DEIX, DISC, TOP
  - Property of proper names: PNLENGTH[5]
- ANTE features:
  - Properties of the antecedent: ANImacy, Clause Type, grammatical FUNction, thematic ROLE, PROXimity, MOOD of the clause in which it is situated.
  - Properties of the relation: REL type (similar to the REF attribute in the MUC scheme, with values IDENT, SUBSET, etc.)

We then ran a preliminary reliability test; this is discussed next.

Reliability

Empirical studies of NP use typically involve a single annotator annotating her corpus according to her own subjective judgment (Prince, 1981; Di Eugenio, 1998; 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.

We studied the reliability of our scheme 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. We discovered that although this was not a truly challenging test of the reliability of our scheme, we nevertheless had significant disagreements in annotating some of the features. The results for the NE attributes were as follows:

Attribute	**K Value**
PNLENGTH	.98
CAT	.92
DISC	.72
LOEB	.63
CB	.6
DEN	.456
TOP	.375

A value of K between .8 and 1 indicates good agreement; a value between .6 and .8 indicates some agreement. Of the features we used, DEN and TOP fail to reach the level of minimum agreement.

Reliability for ANTE features was as follows:

Attribute	**K Value**
ANI	.88
FUN	.68
PROX	.61
REL	.6
CT	.51
ROLE	.42

The disagreements among our annotators had three main causes. In some cases the problem

[4] The element proposed in the MATE scheme derives in turn from the LINK element in the annotation scheme proposed by the Text Encoding Initiative.

[5] For future use - studying when the full form of a proper name is used, as in *John Brown*, or a reduced form, such as *John*.

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 other cases, especially with semantic attributes such as 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. 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 tortoise shell and brass 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 the least reliable feature (TOP), to drastically modify the instructions for annotating the two attributes with next lowest reliability (DEN and LOEB) and to separate out the problem of modification from that of deciding the category by introducing a new attribute MOD. We also completely rewrote the manual. Further revisions of the annotation scheme are foreseen, also to include new information that we discovered was needed (see below); we also plan a second reliability test of the modified scheme.

The Annotation: Methodology, Current Status

We found that annotating both the NE and the ANTE attributes at once was too much work, and led to poor results; and that it was useful to preliminary annotate clauses, as well. We just completed the annotation of the NE features, and are in the process of revising the clause annotation.

4 OUR CORPUS

One of the aims of our project is to verify the generality of our NP generation algorithms by incorporating them in two distinct systems: the ILEX system, that 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;[6] and the ICONOCLAST system, that supports the creation of pharmaceutical leaflets by means of the WYSIWYM technique in which text generation and user input are interleaved (Scott et al., 1998).

The corpus we have collected for GNOME consists of (i) a collection of texts in the museum domain, including most of the corpus collected by the SOLE project, also related to ILEX, to train statistical concept-to-speech algorithms (Hitzeman et al., 1998); and (ii) four texts from the corpus of pharmaceutical leaflets collected for the ICONOCLAST project. Each of the two subsets of our corpus is about 5,000 words and contains about 1,500 NPs.

The texts in the corpus contain examples of all types of NPs 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*). The distribution of various types of NPs in the corpus is shown in the following table.

NP Type	Total Number	Percentage
BARE-NP	700	22.0%
THE-NP	596	18.7%
PN	321	10.1%
PERS-PRO	311	9.80%
A-NP	260	8.19%
POSS-NP	244	7.68%
POSS-PRO	204	6.42%
Q-NP	128	4.03%
COORD-NP	109	3.43%
THIS-NP	88	2.77%
NUMWD-NP	67	2.11%
GERUND	48	1.51%
THIS-PRO	23	0.72%
NUMFIG-NP	23	0.72%
Q-PRO	11	0.34%
ANOTHER-NP	6	0.18%
WH-NP	6	0.18%
NULL-ANA	4	0.12%
REFL-PRO	4	p0.12%
ONE-ANA	3	0.09%
WH-PRO	1	0.03%
THAT-PRO	1	0.03%
THAT-NP	1	0.03%

The range and variety of NPs in the corpus ensures that it will be a rigorous evaluation test-bed of our algorithm's capabilities.

[6] The system is described in (Hitzeman et al., 1997).

Sys cl.:	PersPro	PossPro	TheNP	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	BareNP	Ger	Total	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).

5 TESTING THE SCHEME

We evaluated our scheme by annotating our corpus according to the revised version and using this annotation to build a statistical model of the process of CAT determination. We tried both the Maximum Entropy model (Berger et al., 1996) as implemented by (Mikheev, 1998) and the CART model of decision tree construction (Breiman et al., 1984); the results below were obtained using CART. We did a 10-fold cross-validation.

So far we only completed the annotation of the NE elements. Using only NE features, we obtained a 70% accuracy. (Always choosing the most common category (BARE-NP) results in an accuracy of 22%.) 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, the algorithm gets THIS-NPs, POSS-PROs and GERUNDs mostly wrong, and for the 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 possessive NPs, what is missing is simply the information that the object denoted by the NP is 'owned' by some other entity; this information will be available once we have completed the ANTE annotation, since that will also include information that there is a possession relation between the whole NP and the possessor.[7]

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.

The most complex problem to fix is that of this-NPs: here the problem 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.

An interesting case is that of COORD-NPs. The good results obtained for this class are mainly due to the fact that a number of attributes simply do not apply to this category. It's not clear to us at the moment if we should have a COORD-NPs class at all, or simply treat coordination as a special case of the problem of NP structure (on which we plan to work next).

We observed during the first experiment that the distribution of syntactic categories in the two domains was fairly different. We trained separate models for each domain; the results stayed about the same for the pharmacy domain, whereas we got a better accuracy for the museum domain, 74%.

[7] The set of relations is derived from the one proposed in the MATE scheme.

6 DISCUSSION

It is too early to say whether automatically acquired language models are a practical alternative to hand-coded algorithms in the case of generation; however, our experiments lead us to believe that the methodology we have been using could result in a reasonable performance, once the most obvious problems with the current annotation scheme have been fixed. Annotating a corpus with our scheme is time-consuming, but not excessively so: it took us about 10 man-weeks to complete the annotation of 3,000 NPs, including revisions. This annotation will have to be revised, of course, but once the scheme gets more stable (and the instructions better) we expect that building a statistical model by annotating a corpus and training will not take longer, and may well take less time, than implementing a more traditional algorithm.

It is less clear to us whether such an annotation can really be used to get an evaluation of the performance of more traditional generation algorithms, either to get a feeling for whether the set of features used by such an algorithm is sufficient for the purpose, or maybe even by running the algorithm itself over the annotated corpus (i.e., using the features of each `NE` as input to the algorithm, and comparing the output of the algorithm with the annotated `CAT` value). This is because there is a mismatch 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. First of all, 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, 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. Thus, if the system wants to introduce the new discourse entity *Alphonse Mucha* while at the same time communicating to the user that he was a Czech painter, the system will use a definite description, *the Czech painter Alphonse Mucha*, instead of a proper name. 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.

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, as needed to use the `LOEB` feature. 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).

For all of these reasons, the work described here must only be taken as a first step; we are in the process of thoroughly revising the annotation scheme.

Acknowledgments

The GNOME project is supported by the UK Research Council EPSRC, GR/L51126. Massimo Poesio is supported by an EPSRC Advanced Research Fellowship.

References

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72.

L. Breiman, J. H. Friedman, R. A Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Chapman and Hall.

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

N. A. Chinchor 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*, pages 21–26, Stanford.

R. Dale and E. Reiter. 1995. Computational interpretations of the gricean mixture in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

R. Dale. 1992. *Generating Referring Expressions*. The MIT Press, Cambridge, MA.

S. Davies, M. Poesio, F. Bruneseaux, and L. Romary. 1998. Annotating coreference in dialogues: A proposal for a scheme for MATE. Available at http://www.cogsci.ed.ac.uk/ poesio/MATE/ anno_manual.html, August.

B. Di Eugenio. 1998. Centering in italian. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, chapter 7, pages 115–138. Oxford.

S. C. Garrod and A. J. Sanford. 1983. Topic dependent effects in language processing. In G. B. Flores D'Arcais and R. Jarvella, editors, *The Process of Language Comprehension*, pages 271–295. Wiley, Chichester.

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. (The paper originally appeared as an unpublished manuscript in 1986.).

J. Hitzeman, C. Mellish, and J. Oberlander. 1997. Dynamic generation of museum web pages: The intelligent labelling explorer'. *Journal of Archives and Museum Informatics*, 11:107–115.

J. Hitzeman, 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),*, page Paper 591, Australia.

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

A. Mikheev. 1998. Feature lattices for maximum entropy modeling. In *Proc. of ACL-COLING*, pages 845–848, Montreal, CA.

M. O'Donnell, H. Cheng, and J. Hitzeman. 1998. Integrating referring and informing in NP planning. In *Proc. of the Coling-ACL '98 Workshop on the Computational Treatment of Nominals*, pages 46–55, Université de Montreal, Canada., August.

R. Passonneau and D. Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL.*

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

R. Passonneau. 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, chapter 17, pages 327–358. Oxford University Press.

M. Poesio and R. Stevenson. To appear. *Salience: Computational Models and Psychological Evidence.* Cambridge University Press, Cambridge and New York.

M. Poesio, F. Bruneseaux, S. Davies, and L. Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker, M. Danieli, J. Moore, and B. Di Eugenio, editors, *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging.*

E. F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

D. Scott, 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.

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