A Computational Theory of Prose Style for Natural Language Generation

David D. McDonald and James D. Pustejovsky

Department of Computer and Information Science
University of Massachusetts at Amherst

1. Abstract

In this paper we report on initial research we have conducted on a computational theory of prose style. Our theory speaks to the following major points:

1. Where in the generation process style is taken into account.
2. How a particular prose style is represented; what "stylistic rules" look like;
3. What modifications to a generation algorithm are needed; what the decision is that evaluates stylistic alternatives;
4. What elaborations to the normal description of surface structure are necessary to make it usable as a plan for the text and a reference for these decisions;
5. What kinds of information decisions about style have access to.

Our theory emerged out of design experiments we have made over the past year with our natural language generation system, the Zetalisp program MUMBLE. In the process we have extended MUMBLE through the addition of an additional process that now mediates between content planning and linguistic realization. This new process, which we call "attachment", provides the further significant benefit that text structure is no longer dictated by the structure of the message: the sequential order and dominance relationships of concepts in the message no longer force one form onto the words and phrases in the text. Instead, rhetorical and intentional directives can be interpreted flexibly in the context of the ongoing discourse and stylistic preferences. The text is built up through composition under the direction of linguistic organizing principles, rather than having to follow conceptual principles in lockstep.

We will begin by describing what we mean by prose style and then introducing the generation task that lead us to this theory, the reproduction of short encyclopedia articles on African tribes. We will then use that task to outline the parts of our theory and the operations of the attachment process. Finally we will compare our techniques to the related work of Davey, McKeown and Derr, and Gabriel, and consider some of the possible psycholinguistic hypotheses that it may lead to.

2. Prose Style

Style is an intuitive notion involving the manner in which something is said. It has been more often the professional domain of literary critics and English teachers than linguists, which is entirely reasonable given that it involves optional, often conscious decisions and preferences rather than the unconscious, inviolable rules that linguists term Universal Grammar.

To illustrate what we mean by style, consider the three paragraphs in Figure 1. As we see it, the first two of these have the same style, and the third has a different one.

The Ibibio are a group of six related peoples living in southeastern Nigeria. They have a population estimated at 1,500,000, and speak a language in the Benue-Niger subfamily of the Niger-Congo languages. Most Ibibio are subsistence farmers, but two subgroups are fishermen.

The Ashanti are an AKAN-speaking people of central Ghana and neighboring regions of Togo and Ivory Coast, numbering more than 900,000. They subsist primarily by farming cacao, a major cash crop.

The Ashanti are an African people. They live in central Ghana and neighboring regions of Togo and Ivory Coast. Their population is more than 900,000. They speak the language Akan. They subsist primarily by farming cacao. This is a major cash crop.

Figure 1 Three paragraphs, two styles

The first two of these paragraphs are extracted from the Academic American Encyclopedia; they are the lead paragraphs from the two articles on those respective tribes. The third paragraph was written by taking the same information that we have posited underlies the Ashanti paragraph and regenerating from it with an impoverished set of stylistic rules.

We began looking at texts like these during the summer of 1983, as part of the work on the "Knoesphere Project" at Atari Research (Borning et al [1983]). Our goal in that project was to develop a representation for the kind of information appearing in encyclopedias which would not be tied to the way in which it would be presented. The same knowledge base objects were to be used whether one was recreating an article like the original, or making a simpler version to give to children, or answering isolated questions about the material, or giving an interactive multi-media presentation coordinated with maps and icons, and so on.

With the demise of Atari Research, this ambitious goal has had to be put on the shelf; we have, however, continued to work with the articles on our own. Research on these articles lead us to begin work on prose style. This remains an interesting domain in which to explore style since we are working with a body of texts whose organization is not totally dictated by its internal form.

These paragraphs are representative of all the African tribe articles in the Academic American, which is not surprising since all of the articles were written by the same person and under tight editorial control. What was most striking to us when we first looked at these articles was their similarity to each other, both in the information they contained and the way they were structured as a text. We will assume that for such texts, "encyclopedia style" involves at least the following two generalizations: (1) be consistent in the information that you provide about each tribe; and (2) adopt a complex, "information loaded" sentence structure in your presentation. This sentence structure is typified by a rich set of syntactic constructions, including the use of conjunction reduction, reduced relative clauses, coordination, secondary adjunction, and prenominal modification whenever possible.

A contrasting style might be, for example, one that was aimed at children; we have rewritten the information on the Ashanti tribe as it might look in such a style. We have not yet tried implementing this style since it will call for doing lexicalization under stylistic control, which we have not yet designed.

"The Ashanti are an African people. They live in West Africa in a country called Ghana and in parts of Togo and the Ivory Coast. There are about 900,000 people in this tribe, and they speak a language named AKAN. Most of the Ashanti are cacao farmers."

Figure 2

The style of the Academic American paragraphs, on the other hand, is much tighter, with more compact sentence structure, and a more sophisticated choice of phrasing. Such differences are the sort of thing that rules of prose style must capture.

3. Our Theory of Generation

Looking at the generation process as a whole, we have always presumed that it involved three different stages, with our own research concentrating on the last.

(1) **Determining what goals to (attempt to) accomplish with the utterance.** This initiates the other activities and posts a set of criteria they are to meet, typically information to be conveyed (e.g. pointers to frames in the knowledge base) and speech acts to be carried out.

(2) **Deciding which specific propositions to express and which to leave for the audience to infer on their own.** This cannot be separated from working out what rhetorical constructions to employ in expressing the specified speech acts or from selecting the key lexical items for communicating the propositions. The result of this activity is a *text plan*, which has a principally conceptual vocabulary with rhetorical and lexical annotations. The text plan is seen by the next stage as an executable "specification" that is to be incrementally converted into a text. The specification is given in layers, i.e. not all of the details are planned at once. Later, once the linguistic context of the units within the specification has been determined, this planner will be recursively invoked, unit by unit, until the planning has been done in enough detail that only linguistic problems remain.

(3) **Maintaining a representation of the linguistic structure of the utterance, traversing and interpreting this structure to produce the words of the text and constrain further decisions.** This stage is responsible for the grammaticality of the text and its fluency as a discourse (e.g. insuring that the correct terms are pronominalized, the correct focus maintained, etc.). The central representation is an explicit model of the *surface structure* of the text being produced, which is used both to determine control flow and to constrain the activities of the other processes (see discussion in McDonald [1984]). The surface structure is defined in terms of a stream of phrasal nodes, constituent positions, words, and embedded information units (which will eventually have to be sent back to the planner and then realized linguistically, extending the surface structure in the process). The entities in the stream and their relative order is indelible (i.e. once selected it cannot be changed); however more material can be spliced into the stream at specified points.

3.1 WHERE IS STYLE CONSIDERED?

According to our theory, **prose style is a consequence of what *decisions* are made during the transition from the conceptual representational level to the linguistic level.** The conceptual representation of what is to be said—the text

plan—is modeled as a stream of information units selected by the content planning component. The *attachment process* takes units from this stream and positions them in the surface structure somewhere ahead of the point of speech.

The prose style one adopts dictates what choice the attachment process makes when faced with alternatives in where to position a unit: should one extend a sentence with a nonrestrictive relative clause or start a new one; express modification with an prenominal adjective or a postnominal prepositional phrase. The collective pattern of such decisions is the computational manifestation of one's style.

3.2 EXTENSIONS TO THE SURFACE STRUCTURE REPRESENTATION

The information units from the text plan are positioned at one or another of the predefined "attachment points" in the surface structure. These points are defined on structural grounds by a grammar, and annotated according to the rhetorical uses they can be put to (see later example in Figure 8). They define the grammatically legitimate ways that the surface structure might be extended: another adjective added to a certain noun phrase, a temporal adjunct added to a clause, another sentence added to a paragraph, and so on.

Which attachment points exist at any moment is a function of the surface structure's configuration at that moment and where the point of speech is. Since the configuration changes as units are added to the surface structure or already positioned units are realized, the set of available attachment points changes as well. This is accomplished by including the points in the definitions of the phrasal elements from which the surface structure is built. We have since argued that this addition of attachment point specifications to elementary trees is very similar to the grammatical formalism used in Tree Adjoining Grammars [Joshi 1983] and are actively exploring the relationships between the two theories (cf. McDonald & Pustejovsky [1985a].)

3.3 A DECISION PROCEDURE

The job of the attachment process is to decide which of the available attachment points it should use in positioning a text plan unit in the surface structure. This decision is a function of three kinds of things:

1. The different ways that the unit can be realized in English, e.g. most adjectives can also be couched as relative clauses, not all full clauses can be reduced to participial adjectives.
2. The characteristics of the available attachment points, especially the grammatical constraints that they would impose on the realization of any unit using them. The "new sentence" attachment will require that the unit be expressible as a clause and rule out one that could only be realized as a noun phrase; attachment as the head of a noun phrase would impose just the opposite constraint.
3. What stylistic rules have been defined and the predicates they apply to determine their applicability.

The algorithm goes as follows. The units in the stream from the text plan are considered one at a time in the order that they appear. There is no buffering of unpositioned units and no lookahead down the stream to look for patterns among the units; any patterns that might be significant are supposed to already have been seen by the text planner and indicated by passing down composite units.[1] Each unit is thus considered on its own, on the basis of how it can be realized.

The total set of alternative phrasings for an information unit are precomputed and stored within the linguistic component (i.e the third stage of the process) as a "realization class". Different choices of syntactic arrangement, optional arguments, idiomatic wordings, etc. are anticipated before hand (by the linguist, not the program) and grouped together along with characteristics that describe the uses to which the different choices can be put: which choice focuses which argument; which one presumes that the audience will already understand a certain relationship, which one not. (Realization classes are discussed at greater length in McDonald & Pustejovsky [1985b]).

The first step in the attachment algorithm is to compute all legitimate pairings of attachment points and choices in the unit's realization class, e.g. a unit might be attached at a NP premodifier point using its adjective realization; or as postmodifier using its participial realization; or as the next sentence in the paragraph using any of its several realizations as a root clause. This particular case is the one in our example in Section 4.

The characteristics on each of the active attachment points will be compared with the characteristics on each of the choices in the unit's realization class. Any choice that is compatible with a given attachment point is grouped with it in a set; if that attachment point is selected, a later decision will be made among the choices in that set.

Once the attachment point/choice set pairs have been computed, the next step is to order them according to which is most consistent with the present prose style. This is where the stylistic rules are employed. Once the pairs are ordered, we select the pair judged to be the best and use it. The unit is spliced into the surface structure at the selected attachment point, and the choices consistent with

---

1 Assuming that the criterial division between conceptual/rhetorical planning and linguistic realization is that only the linguistic side understands grammar, e.g. the opportunities and constraints implicit in the surface structure at a given moment (we think that both sides should be designed to appreciate the lexicon), then this restriction implies that there will be no opportunistic reconfiguring of the text plan by the linguistic component, no condensing parallel predicates into conjunctions or grouping of modifiers etc. unless there is a specifically planned rhetorical motive for doing so dictated by the planner.

that point set up for later selection (realization of the unit) once that point is reached by the linguistic component in its traversal.

3.4 STYLISTIC RULES

As we have just said, the computational job of a stylistic rule is to identify preferences among attachment points.[2] This means that the rules themselves can have a very simple structure. Each rule has the following three parts:

1. A name. This symbol is for the convenience of the human designer; it does not take part in the computation.
2. An ordered list of attachment points.
3. A predicate that can be evaluated in the environment accessible within the attachment process. If the predicate is satisfied, the rule is applicable.

Each stylistic rule states a preference between specific attachment points, as given by the ordering it defines. To perform the sorting then, one performs a fairly simple calculation (n.b. it is simple but lengthy; see footnote).

(1) For each candidate attachment point, collect all of the stylistic rules that mention it in their ordered lists; discard any rules that do not mention at least one of the other candidate points as well.

(2) Evaluate the applicability predicates of the collected rules and discard any that fail.

(3) Using the rules that remain, sort the list of candidate attachment points so that its order matches the partial orders defined by the individual stylistic rules.

We have now looked at our treatment of four of the five points which we said at the onset of this paper had to be considered by any theory of prose style. The fifth point, the kinds of information stylistic rules are allowed to have access to, requires some background illustration before it can be addressed; we will take it up at the end of our example.

4. An Example

4.1 Underlying representation

At the present time we are representing the information about a tribe in a frame language known as ARLO [Haase 1984], which is a CommonLisp implementation of RLL. We have no stock in this representation per se, nor, for that matter, in the specific details of the frames we have built (though we are fairly pleased with both); our system has worked from other representations in the past and we expect to work with still others in the future. Rather, this choice provides us with an expeditious, non-linguistic source for the articles, which has the characteristics we expect of modern representations. Figure 2 shows the toplevel ARLO frame for the Ashanti and one of its subframes.

```
(defunit Ashanti
   (Prototype #>african-tribe)
   (encyclopedia-unit? t)
   (location #>Ashanti-location)
   (population #>Ashanti-population)
   (language #>Akan)
   (economic-bases #>Ashanti-economy))

(defunit #>Akan
   (Prototype #>language
   (encyclopedia-unit? t)
   (speaker #>Ashanti))
```

Figure 3 Ashanti ARLO-unit

Given this representation, it is a straightforward matter to define a fixed script that can serve as the message-level source for the paragraphs. We simply list the slots that contain the desired information.[3]

```
(define-script  default-tribe-introductory-paragraph  (tribe)
   ( #>prototype
     #>language
     #>alternative-names
     #>location
     #>population
     #>economic-basis
     #>distinguishing-physical-characteristics) )
```

Figure 4 Text Script Structure

[2] At present "preference" is defined by sorting candidate point–choice pairs against the rules and selecting the topmost one; it is easy to see how less computationally intensive schemes could be worked out. Some stylistic rules should probably be allowed to "veto" whole classes of attachment points and others able to declare themselves always the best. Furthermore these rules naturally fall into groups by specialization and features held in common, suggesting that the "sort" operation could be sped up by taking advantage of that structure in the algorithm rather than simply sorting against all of the stylistic rules uniformly. We have worked out on paper how such alternatives would go, and expect to implement them later this year.

[3] In ARLO slots are first-class objects with a prototype hierarchy of their own just like the one for units (frames). The list of slots is effectively a list of access functions whose domain is units (the tribe being described) and whose range is also units (the slot values). When this script is instantiated, the generator will receive a list of triple records: slot, unit, and value.

If any of these slots are empty or "not interesting" for the tribe, it is simply left out. The interface between planner and realization can be this simple because the type of text we are generating is fairly programmatic and predictable. With a more complicated task comes a more sophisticated planner. The point here, however, is to examine a simple planning domain in order to isolate those decisions that are purely stylistic in nature.

4.2 Attachment

To illustrate what attachment adds, let us first look what the usual alternative procedure, direct translation,[4] would do with the information plan we use for these paragraphs. It would realize the items in the script one by one, maintaining the given order, and the resulting text would look like this (assuming the system had a reasonable command of pronominalization):

The Ashanti are an African people. They live in central Ghana and neighboring regions of Togo and Ivory Coast. This is in West Africa. Their population is more than 900,000. They speak the language Akan. They subsist primarily by farming cacao. This is a major cash crop.

Figure 5 Paragraph II by *Direct Replacement*

Although true to the information in the script, this method does not reflect the complex stylistic variations and enrichments that make up the original paragraph. There must be something above the level of a single information unit to coordinate the flow of text, while not altering the intentions or goals of the planner. With this in mind, we have built a stylistic controller which has the following properties:

- o It allows information to be "folded in" to already planned text. Items in the script do not necessarily appear in the same order in the text.
- o The decision about when to fold things in is made on the basis of style; i.e. if the style had been different, the text would have been different as well.
- o The points where new material may be added to planned text are defined on structural grounds.

For example, notice that in paragraph II from Figure I the language-field is realized as as a compound adjectival phrase, modifying the prototype; viz. "Akan-speaking." For the first article, however, the language-field is realized differently. The attachment-point that allows this "fold-in" (i.e. attach-as-adjective) is introduced by the realization class for the prototype field. The decision to select this phrase over the sentential form in Figure 5 is made by a **stylistic rule.** This rule (cf. Figure 6) states that the adjectival form is preferred if the language name has its own encyclopedia entry.[5] We see that this stylistic rule is not satisfied in Paragraph I, hence another avenue must be taken (namely, clausal). The other attachment points used by the stylistic rules determine whether to use a reduced relative clause, a new sentence, or perhaps an ellipsed phrase. The stylistic rule allowing this structure is given below in Figure 6.

```
(define-stylistic-rule  PREFER-NOUN-ADJ-COMPOUND-TO-POSTNOM
  ordering-on-attachment-points
     ( attach-as-adjective  attach-as-postnominal-phrase )
  applicability-condition
     (encyclopedia-entry? Noun) )

(define-stylistic-rule  PREFER-ADJECTIVES-TO-NEW-SENTENCE
  ordering-on-attachment-points
     (attach-as-adjective attach-as-new-sentence )
  applicability-condition
   (if (includes-attachment-point 'attach-as-adjective
         usable-attachment-points)
     (not (or (will-be-complex-adjective-phrase
                 (usable-choices 'attach-as-adjective))
              (too-heavy-with-adjectives
                (np-being-attached-to 'attach-as-adjective))))))
```

Figure 6 Stylistic Rules

Consider now the derivation of the first sentence of Paragraph I, and how the stylistic rules constrain the attachment process. The first unit to be planned as surface structure is the prototype field—the essential attribute of the object. This introduces, as mentioned above, an attachment point on the NP node, allowing additional information to be added to the surface structure. The realization class associated with the language field for the Ashanti is transitive-verb, represented in Figure 7 below.

---

[4] "Direct translation" is a term coined by Mann et al. [1981] to describe the techniques used by most of the generation systems in use to day with working expert systems. It entails taking a complex structure from the system's knowledge base as the text source (in this case our list of slots) and building from it a discourse that matches it exactly in structure by recursively selecting texts for its sourse.

[5] This rule is particular to the encyclopedia domain, of course, and makes reference to information specifically germaine to encyclopedias. The rule, however, is to the point, and appears to be productive; e.g. "wheat farmers", "town dwellers", etc.

```
(define-realization-class  transitive-verb
  : parameters  (agent object verb)
  : choices
     ((  (default-active-form verb agent object)
         clause)
   ; A speaks B
      (  (passive-form verb agent object)
         clause in-focus(obj) )
   ; B is spoken by A
      (  (gerundive-with-subject verb subj obj)
   ; A speaking B
         np)
      (  (gerundive-passive-with-subject verb subj obj)
   ; B being spoken by A
         np in-focus(obj))
      (  (adjectival-form verb object)
         AdjP expresses-theme(B) )
      : B-speaking
  )
```

Figure 7 Realization Class for Transitive Verb

Because of the stylistic rules, the compound-adjectival form is preferred. The preconditions are satisfied—namely, *Akan* is itself an entry in the encyclopedia— and the attachment is made. Figure 8 shows the structure at the point of attachment.

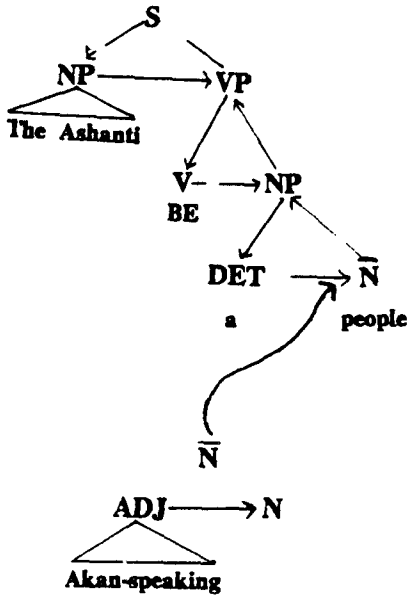

Figure 8 Attachment of (language #<Akan)

5. Comparisons with other Research in Language Generation

Two earlier projects are quite close to our own though for complementary reasons. Derr and McKeown [1984] produce paragraph length texts by combining individual information units of comparable complexity to our own, into a series of compound sentences interspersed with rhetorical connectives. Their system is an improvement over that of Davey [1978] (which it otherwise closely resembles) because of its sensitivity to discourse-level influences such as focus.

The standard technique for combining a sequence of conceptual units into a text has been "direct replacement" (see discussion in Mann et al. [1982]), in which the sequential organization of the text is identical to that of the message because the message is used directly as a template. Our use of attachment dramatically improves on this technique by relieving the message planner of any need to know how to organize a surface structure, letting it rely instead on explicitly stated stylistic criteria operating after the planning is completed.

Derr and McKeown [1984] also improve on direct replacement's one-proposition-for-one-sentence forced style by permitting the combination of individual information units (of comparable complexity to our own) into compound sentences interspersed with rhetorical connectives. They were, however, limited to extending sentences only at their ends, while our attachment process can add units at any grammatically licit position ahead of the point of speech. Furthermore they do not yet express combination criteria as explicit, separable rules.

Dick Gabriel's program *Yh* [1984] produced polished written texts through the use of critics and repeated editing. It maintained a very similar model to our own of how a text's structure can be elaborated, and produced texts of quite high fluency. We differ from Gabriel in trying to achieve fluency in a single online pass in the manner of a person talking off the top of his head; this requires us to put much more of the responsibility for fluency in the pre-linguistic text planner, which is undoubtedly subject to limitations.

It is our belief that, for script-like domains, online text generation suffices. This method, in fact, provides us with an interesting diagnostic to test our theory of style: namely, that stylistic rules are meaning-preserving, and do not change the goals or intentions of the speaker. Stylistic rules are to be distinguished from those syntactic rules of grammar which affect the semantic interpretation of a syntactic expression. A non-restrictive relative, for example, is a particular stylistic construction that adds no meaning-delimiting predication to the denotation of the NP. Use of a restrictive relative, on the other hand, is not a matter of style, but of interpretation; "the man who owns a donkey" is not a stylistic variant of the proposition "The man owns a donkey." In other words, the stylistic component has no reference to intentions, goals, focus, etc.

These are the concerns of the planner, and are expressed in its choices of information units and their description (cf. Mann and Moore [1983] for a discussion of similar concerns).

6. Status and Future Work: Computational Models of Text Planning

At the time this is being written, the core data structures and interpreters of the program have been implemented and debugged, along with the set of attachment-points and stylistic rules, which are necessary to reproduce the paragraphs. The stylistic planner is completely integrated with the language generation program and has produced texts for scene descriptions (McDonald and Conklin (forthcoming)), narrative summaries (Cook, Lehnert, McDonald, [1984]), and two of the three paragraphs shown in Figure 1.

Currently we are shifting domains to generate newspaper articles, in the style of the New York Times. We have only a single style worked out in detail, but we would like to handle styles involving alternative lexical choices, as well.

Ultimately what is most exciting to us is the opportunity that we now have to use this framework to develop precise hypotheses about the nature of the "planning unit" in human language generation. This has been an important question in psycholinguistic research as well (Garrett [1982]). This continues our ongoing line of research on the psychological consequences of our computational analysis of generation. The following are a few of the questions that must be addressed in the research on planning:

- What is the size of the planning units at various stages;
- What is the vocabulary that the units are stated in, e.g. are conceptual and linguistic objects mixed together or are there distinct unit-types at different levels, with some means of cascading between levels;
- Should units be modelled as "streams" with conceptual components passing in at one end and text passing out at the other, or are they "quanta" that must be processed in their entirety one after the other; and finally
- Can the components of a planning unit be revised after they are selected, or may they only be refined. This appears to relate to similar questions in psycholinguistic research (see Garrett [1982] for review).

7. Acknowledgements

This research has been supterminaled in part by contract N0014-85-K-0017 from the Defense Advanced Research Projects Agency. We would like to thank Marie Vaughan for help in the preparation of this text.

8. References

Borning, A., D. Lenat, D. McDonald, C. Taylor, & S. Weyer (1983) "Knoesphere: Building Expert Systems with Encyclopedic Knowledge" Proc. IJCAI-83, pp.167-169.

Cook, M., W. Lehnert, & D. McDonald (1984) "Conveying Implicit Context in Narrative Summaries", Proc. of COLING-84, Stanford University, pp.5-7.

Davey (1974) *Discourse Production*, Ph.D. Dissertation, Edinburgh University; published in 1979 by Edinburgh University Press.

Derr,M. & K. McKeown (1984) "Using Focus to Generate Complex and Simple Sentences" Proceedings of COLING-84, pp.319-326.

Gabriel R., (184) Ph.D. thesis, Computer Science Department, Stanford University.

Gabriel, R. (forthcoming) "Deliberate Writing" in Bolc (ed.).

Garrett, M. (1982) "Production of Speech: Observations from Normal and Pathological Language Use", in *Pathology in Cognitive Functions*, London, Academic Press.

Haase, K. (1984) "Another Representation Language Offer", Ph.D. Thesis, MIT.

McDonald,D. (1984) "Description Directed Control: Its implications for natural language generation", *International Journal of Computers and Mathematics*, 9(1) Spring 1984.

McDonald,D. & E. J. Conklin (in preparation) "At the Interface of Planning and Realization" in Bloc and McDonald (eds.) *Natu l Language Generation Systems*, Springer-Verlag.

McDonald D., & Pustejovsky J. (1985a) "TAGs as a Grammatical Formalism for Generation", Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, University of Chicago.

McDonald D. & Pustejovsky J. (1985b) "Description-Directed Natural Language Generation", Proceedings of IJCAI-85, W.Kaufmann Inc., Los Altos CA.

Mann W., Bates M., Grosz G., McDonald D., McKeown K., Swartout W., "Report of the Panel on Text Generation" Proceedings of the Workshop on Applied Computational Linguistics in Perspective, American Journal of Computational Linguistics, 8(2), pgs 62-70.