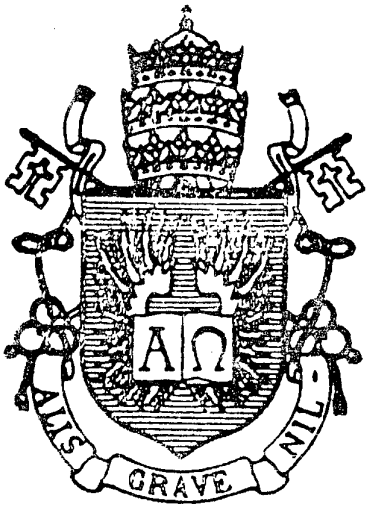

PUC

Series: Monografias em Ciência da Computação, 8/89

ENHANCING TEXT QUALITY IN QUESTION ANSWERING SYSTEMS

Clarisse S. Souza
Donia R. Scott
Maria das Graças V. Nunes

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
RUA MARQUÊS DE SÃO VICENTE, 225 – CEP 22453
RIO DE JANEIRO – BRASIL

PUC/RJ - DEPARTAMENTO DE INFORMATICA

Series: Monografias em Ciência da Computação, 8/89

Editor: Paulo Augusto Silva Veloso May, 1989

ENHANCING TEXT QUALITY IN QUESTION ANSWERING SYSTEMS

Clarisse S. de Souza
Donia R. Scott
Maria das Graças V. Nunes

This work has been partially sponsored by FINEP

arge of publications:

Isane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC RIO, Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22453 - Rio de Janeiro, RJ
BRASIL

Tel.: (021)529-9386 TELEX:31078 FAX: (021)274-4546
BITNET: userrtlc@lncc.bitnet

hancing Text Quality in a Question-Answering System

Clarisse Sieckenius de Souza [§]

Donia R. Scott [†]

Maria das Graças Volpe Nunes [‡ §]

[§] Dept. de Informática, PUC/RJ
R. Marquês de São Vicente, 225,
22453 Rio de Janeiro / RJ - Brasil

[†] Philips Research Laboratories,
Cross Oak Lane, Redhill,
Surrey RH1 5HA, U.K.

[‡] ICMSC - Universidade de Sao Paulo,
R. Carlos Botelho, 1465,
13560 - São Carlos / SP - Brasil

1. Introduction

Cooperativity has been extensively declared to be a necessary feature of intelligent interactive systems. A cooperative response is one which optimally achieves the responder's communicative goal, which is to change the questioner's mental state from one of not knowing to knowing the facts questioned [1]. To achieve this goal in an optimal way, the responder should not only provide the appropriate information, but do so in a manner which ensures that it will be easily rstood. Leaving aside issues of information content, it is clear that ammaticality is a necessary but not sufficient criterion in this context for judging the quality of a response. We will assume for the sake of discussion that in conversational systems which involve text output, text that is 'good' according to the writing conventions of the language in question is adequate for the situation.

This paper describes some methods of enhancing the textual quality of responses generated within the framework of Rhetorical Structure Theory [2,3], a framework that is currently being explored by the authors for producing textual responses in Brazilian Portuguese [4,5,6] and by others for English [7]. It is clear that although there are general criteria for good text that hold across

languages, fine-grained tuning is required for each language. Our methodology addresses the general problem of producing good-quality textual responses, with particular reference to Brazilian Portuguese.

Rhetorical Structure Theory (RST) is a theory of text organization in which spans of text are described according to the rhetorical relations that hold between them. Elements of a relation are referred to as nuclei and satellites, with nuclei being more semantically primary to the text than satellites. Relations hold between elements of a text at a number of levels (ie. a RST structure is essentially hierarchical). Text coherence is defined by the existence of a rhetorical relation between each part of the text and at least one of its neighbours, and cohesion as the existence of a relation that holds over the entire text.[1]

Each RST relation is defined by the constraints that hold on the content of its elements and on their combination and by a statement of what its impact should be on the reader. This characteristic is especially important in question-answering systems, where beliefs play a major role in the design of cooperative responses.

Although RST is essentially a descriptive theory of text structure, it is particularly attractive for text generation since it encompasses criteria by which a piece of text can be judged as good on both linguistic and cognitive grounds. In a generative process, however, the step between the planning and realisation phases must involve some sort of linguistic strategy for the mapping of structure onto text. Such a strategy would direct the syntactic and lexical choices available for expressing the relations that hold between the information units that form the terminal elements of the structure. Previous research [8] has pointed to what some of these mapping elements would look like.

LETTERA is a program that uses RST as a basis for generating textual responses in Brazilian Portuguese to Yes/No questions about crimes and criminology [5]. The input to LETTERA is an RST schema, the construction of which is guided by focus considerations [4,6]. These schemas keep intact the original specifications of RST relations. The generation of text involves a mapping between RST relations and linguistic markers, and the use of a generative grammar of Portuguese. The grammar operates in two steps: (a) generating basic sentences from clause-sized knowledge-representation structures

---

1. The JOINT schema is an exception to this.

(first-order predicates) and (b) performing transformations on these to produce the final text. A typical output of LETTERA would be:

Question: Pedro foi ferido?
Was Peter hurt?

LETTERA:
Sim, com facadas. O autor do crime é desconhecido e fugiu.
O local do crime é o Leme e a data é dia 21 de junho.
Yes, with stabs. The author of the crime is unknown and has fled. The location of the crime is Leme and the date is June 21.[2]

Although correct in essence, this response is stylistically inadequate. A much better text in Brazilian Portuguese would be:

LETTERA:
Sim, com facadas. O autor do crime, ocorrido no Leme no dia 21 de junho, é desconhecido e fugiu.
Yes, with stabs. The author of the crime, which occurred in Leme on June 21st, is unknown and has fled.

e the second sentence contains an embedded clause. It is indisputable that the meaning of the first text is absolutely clear, the second is much more acceptable as 'fluent' Brazilian Portuguese. A similar problem, which Hovy [7] refers to as one of sentence scoping, occurs in PENMAN.

In the rest of this paper, we describe the technique we have developed for improving the quality of text generated by LETTERA so that it is more like the second than the first example.

2. Expanding RST-based Planning Capabilities

The responses LETTERA produces to Yes/No questions all conform to one of the general schemas shown in Figures 1 and 2. The organisation of these guided by focus considerations and by general principles of cooperativity. n in the figures, degrees of cooperativity vary as one moves through the egions of a schema. Each node in the schema is a relation, with the heavy branches pointing to Nuclei and the light ones to Satellites. The variables x and y correspond to the focussed elements of the reponse; p_i and p_i' correspond to predicates, where p_i' is related, but not equivalent, to p_i.

[2] The English translations given throughout are more-or-less literal ones.

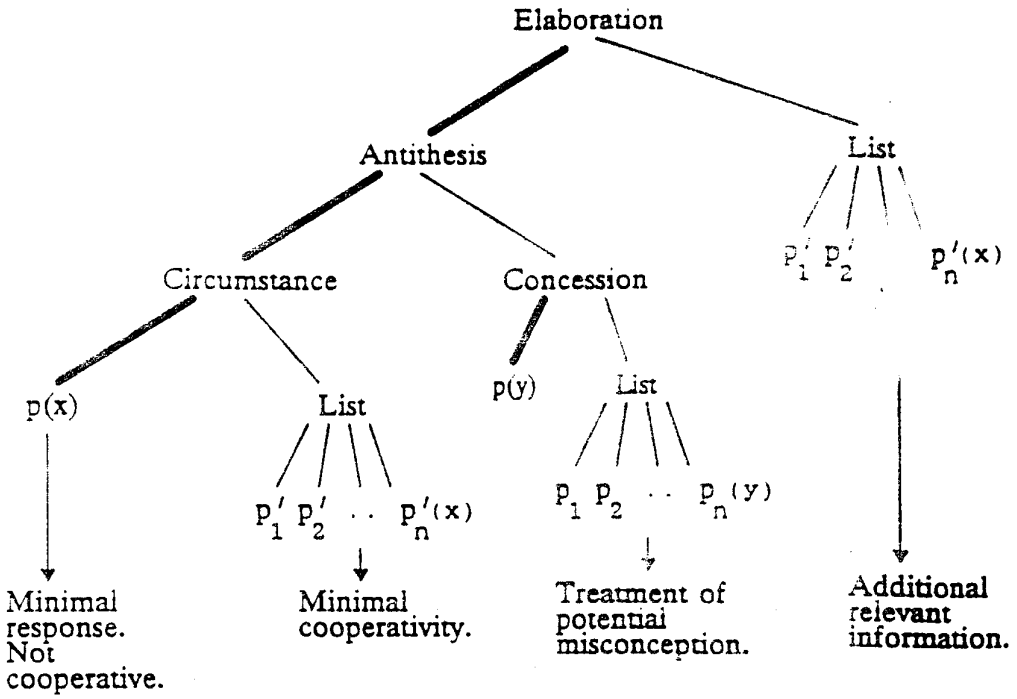

Figure 1: Yes Schema

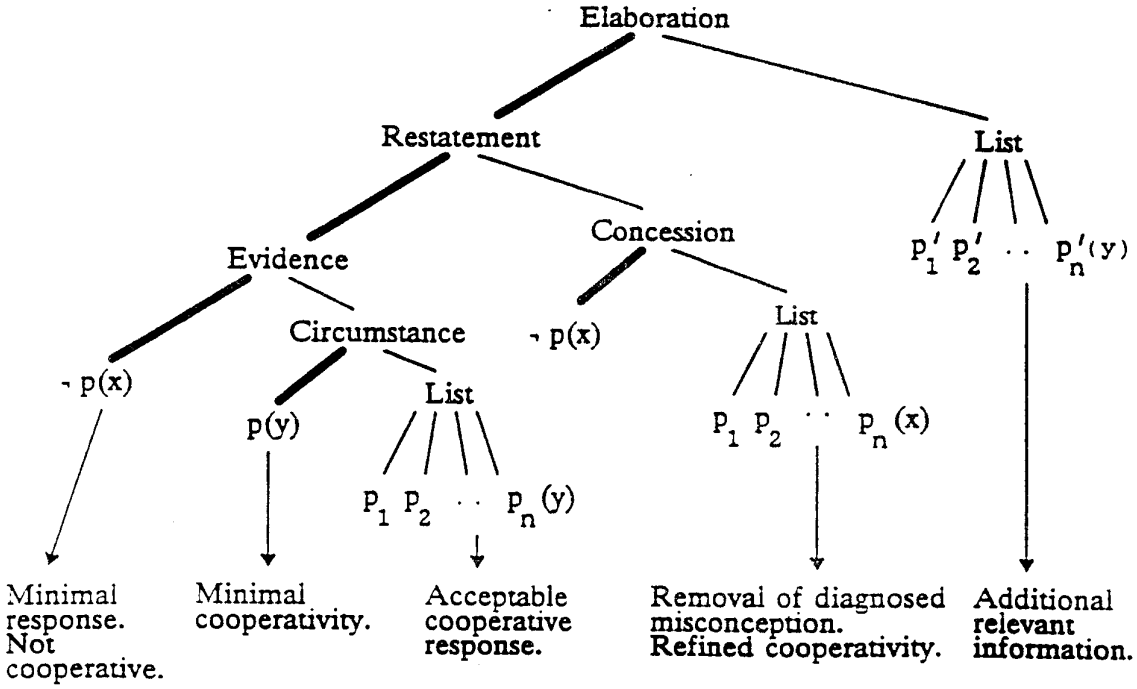

Figure 2: No Schema

Response schemas represent the upper limit on what a response can contain. R nding to Yes/No questions, the output of the planner (the input to LE TERA) is one of the above schemas to which terminal clause-sized elements (first-order predicates) have been added. Realising these schema by traversing them in a depth-first manner, transforming each terminal element into an S-bar with accompanying lexical markers of relations and with anaphoric reference could result in text such as: [3]

Example 1: Yes Schema

Question: Pedro foi atacado?
Was Pedro attacked?

Response:

1. {p(x)} Sim.
   Yes.
2. {p1(x)} Pedro foi atacado com um guarda-chuva.
   Pedro was attacked with an umbrella.
3. {p2(x)} Ele foi atacado com uma bolsa.
   He was attacked with a handbag.
4. {p(y)} Não foi Paulo a vítima,
   Paulo wasn't the victim,
5. {p1(y)} embora ele tivesse roubado a carteira de um pedestre.
   although he had stolen a pedestrian's wallet.
6. {p1'(x)} Pedro costumava roubar as pessoas na rua.
   Pedro is known to have robbed people on the street.
7. {p2'(x)} Ele tinha estado preso várias vezes.
   He has been imprisoned many times.
8. {p3'(x)} Ele tinha sido condenado por pequenos furtos.
   He has been convicted for petty theft.

Example 2: No Schema

uestion: José foi suspeito de ter ferido Pedro?
Was Jose suspected of having injured Pedro?

Response:

1. { ¬p(x)} Não.
   No.
2. {p(y)} José foi testemunha do crime,
   Jose was a witness to the crime,
3. {p1(y)} Uma vez que ele viu o crime.
   since he saw the crime take place.
4. { ¬p(x)} Ao passo que José não foi suspeito,
   Even though Jose was not suspected,
5. {p1(x)} se bem que ele esteve no Leme no dia 21.
   even so he was in Leme on the 21st.

[3] Information on how the content of responses is determined is given in [4] and [6].

6. {p1'(y)} José era primo de Pedro.
Jose was Pedro's cousin.

For purposes of explication, we show the focus tag and number associated with each clause. Of course, the final output of the generator will be rather better presented.

It is fairly apparent from these examples that the set of transformational operations used to produce them is not sufficient to render a stylistically pleasing text. An obvious solution would be expanding the realisation component to allow it to perform more sophisticated syntactic operations for clause-combining. It is clear, however, that such enhancements must be sensitive to the structural context in which terminal elements occur and not simply to the immediate (low-level) context of the neighbouring elements.

For example, from the point of view of style, clause 6 in Example 2 should be embedded in clause 2 but not in clauses 3, 4 or 5. Similarly, it would make good stylistic sense for embedding to operate between clauses 6 and 2, but not coordination. Example 2 also indicates the types of stylistic blunders that can occur when the realiser does not have access to information about the constraints that apply to the markers of a relation: those used for clauses 4 and 5, although appropriate for the relation, should not occur together.

In addition to structure-sensitive syntactic realisation rules, some sort of interaction with linguistic information will be desirable at the planning stage. That is, the rhetorical relations, the elements on which the planner operates to produce the RST structure, must contain as part of their definition a specification of the linguistic operations that can be applied to them.

3. Adding Linguistic Information to RST Relations

We propose that the specifications of RST relations should include the information outlined below. The items presented here are not intended to comprise the full set of information required for the production of stylistically adequate text. Rather, they comprise the minimal set required to produce stylistic enhancements of the type discussed above.

1. **A specification of the permissible syntactic structures in which the relation's Nuclei and Satellites can be linguistically realised.**

The default syntactic realisation of all structure-terminal elements is a sentence (S-bar). However, whereas Nuclei are only realisable as S-bars, Satellites may

be realised as sub-sentential structures (e.g. adjectives, noun-phrases or prepositional phrases). In this respect we depart from the clause-combining proposal of [8], which suggests that satellites should also be realised as S-bars. Instead, we take the view that the semantic subordination of Satellite to Nucleus should be expressible syntactically as embedding. This point is discussed in greater detail below.

2. A specification of the lexical or phrasal markers that apply to the relation, under what conditions and with attachment to which element.

Although the rhetorical relations between two text spans are sometimes retrievable from the mere juxtaposition of the two, this is often not the case. When it is not, then explicit signals of the relation must be given. The more syntax and semantics interact to produce the meaning of a relation, the greater the need will be to explicitly mark that relation.

3. A specification of the permissible permutations of a Nucleus and a Satellite, and the conditions under which they may occur.

As a general principle of RST, there are no constraints on the order in which the elements of a relation can occur. However, a 'Nucleus before Satellite' order is most prevalent in natural text, presumably due to cognitively-motivated factors related to the primacy of Nuclei over Satellites to the text. A 'Satellite before Nucleus' order tends to coincide with the presence of explicit lexical or phrasal markers of the semantic subordination of Satellite to Nucleus.

4. Structure-sensitive Syntactic Rules

Rhetorical relations can be signalled syntactically through subordination and coordination. In general, syntactic subordination reflects the presence of semantic subordination of Satellite to Nucleus, whereas coordination reflects the linking of independent elements — that is, Satellites with other Satellites and Nuclei with other Nuclei (see e.g. [9]). Exceptions to this rule involve cases where coordination reflects semantic subordination, and where subordination is temporal or causal. For example:

> Eles não estão se dando bem, e ela decidiu sair de casa.
> *They're not getting along, and she's decided to move out.*

For reasons of simplicity, we have chosen to use the general rule for making the choice between applying syntactic subordination or coordination and to deal with the exceptions within the specification of relations by treating the coordinating conjunction as a lexical marker. For example, "and" will appear as a lexical marker in the specification of SEQUENCE and relations in the CAUSE cluster.

4.1 Embedding Environments

The semantic subordination of Satellite to Nucleus will be syntactically marked by embedding. Although research on text analysis using RST suggests that embedding will be undesirable, or at the very least that its frequency should be restricted [8], our experience has shown this not to be the case. LETTERA presents a number of situations which favour embedding, most noticeably when the response contains information that is supplementary or complementary to the main idea that it is attempting to convey to the reader. These situations tend to involve the ELABORATION, CONCESSION and CIRCUMSTANCE relations. Given that LETTERA uses only a reduced set of RST relations, it may well be the case that general embedding rules involving all 20 RST relations will be required, but this is outside the scope of the present paper.

Rule 1: A Satellite can only be embedded in its Nucleus.

This restiction on which of the Nuclei of a schema can be a candidate home for an embedded Satellite ensures that embedding does not disturb the hierarchical relationships of the RST structure.

Rule 2: Embedding can be realised as an adjective, appositive noun phrase (*predicativo*), prepositional phrase or relative clause and should be realised in that order of preference.

This rule provides for subordinate structures that do not impair the style of the main clause. Although psycholinguistically motivated, the specification of preferred constructions is to some extent language dependent. The order given in the rule is that which is most appropriate to Brazilian Portuguese, and will require fine-tuning when applied to another language.

Rule 3: Embedding can occur within the left-most nuclear clause in the structure bearing the same focus value as the candidate clause.

Exceptions to the rule will be expressed as syntactic constraints in the specification of relations. For example, Satellite elements of CIRCUMSTANCE

and CONCESSION can only be embedded in their immediate Nucleus, and embedding cannot occur across Nucleus and Satellite of RESTATEMENT.

To maximize stylistic effects, these three rules are applied in the order in which they are presented here. In addition to them, another more global structure-related embedding rule is required:

Rule 4: Satellites in a LIST schema of ELABORATION should (where possible) be embedded, provided that the number of remaining clauses is 0 or greater than 1.

This rule prevents the appearance of 'dangling' sentences in the text. ELABORATION is perhaps the weakest relation, in that the content of its Satellite is less strongly related to that of the Nucleus that in other relations. When there is only one clause in the Satellite of ELABORATION, the effect on the reader is of it being included as an after-thought. This effect increases in relation to increases in the size of the Nucleus and decreases in the size of the Satellite.

4.2 Coordination Environments

As mentioned before, coordination will be applied as a syntactic marker of the absence of semantic subordination. It appears to be the case that, at least for Brazilian Portuguese and English, elements of a text that bear different rhetorical relations to the rest of the text are not suitable candidates for coordination.

The only structural configurations in RST which do not involve coordination are multi-nuclear schemas (CONTRAST, JOINT and SEQUENCE) and multi-satellite ones (LIST and MOTIVATION/ ENABLEMENT). In the light of the above constraints on coordination, only three of these support coordination, leading to the rule:

Rule 5: Coordination can only occur between elements of LIST, SEQUENCE and CONTRAST. [4]

There are a number of other aspects of coordination which affect style and which are rather more related to psycholinguistic than structural factors. These lead to two more global coordination rules which reflect psycholinguistic evidence [10] on the facilitation effects of syntactic factors on sentence processing.

---

[4] JOINT is not included here since it does not lead to coherence.

Rule 6: The greater the number of shared parameter values between clauses, the more desirable it is to coordinate them.

Rule 7: It is more desirable to coordinate Noun Phrases before Prepositional Phrases, which in turn are more desirable candidates for coordination than Verbs or Verb Phrases.

4.3 Heuristics for Sentence Complexity and Rule Ordering

In addition to the above structure-related rules for combining terminal clauses, there are a number of more general ones that are required for the production of stylistically good text. These rules guarantee that the filtering of the number of embedded and coordinated clauses, and the number of levels of embedding, does not degrade the clarity of the derived text. They are:

Rule 8: Sentences should contain no more than 3 clauses, including embedded ones.

Rule 9: Sentences should contain no more than 1 level of embedding.

Rule 10: Embedding should occur before Coordination.

Rule 11: Embedding should occur before focus transformations.

Rule 12: Clauses with time and place predicates should not be coordinated.

These rules are the result of applying native intuitions on the effect of sentence complexity on style (8 and 9), and of the effect of implementations of partial algorithms in LETTERA of the embedding and coordination rules discussed above (10 and 11). Finally, cognitive factors appear to favour time and place as a single predicate when related to the same event. Exceptions to this appear to occur only for reasons of emphatic stress for argumentative purposes.

5. Discussion:

This paper addresses the problems that arise when text is generated by realising the terminal elements of a hierarchical plan in a strictly bottom-up way. It argues, as we have argued elsewhere [11], that the generation of good quality text can only be achieved if the realisation process is sensitive to the structure of the plan and to psycholinguistic factors. We have suggested here some methods of achieving this within the framework of Rhetorical Structure Theory. Applying them to the text plan for Examples 1 and 2 above, we now get:

E ample 1:

Sim. Pedro, que costumava roubar as pessoas na rua, foi atacado com um ıarda-chuva e uma bolsa. Não foi Paulo a vítima, embora ele tivesse ubado a carteira de um pedestre. Pedro tinha estado preso várias vezes, e sido condenado por pequenos furtos.

Yes. Pedro, who is known to have robbed people on the street, was attacked with an umbrella and a handbag. Paulo wasn't the victim, although he is known to have stolen a pedestrian's wallet. Pedro has been imprisoned several times, and has been convicted for petty theft.

Example 2:

Não. José, primo de Pedro, foi testemunha do crime, uma vez que ele viu o crime. José não foi suspeito, embora tivesse estado no Leme no dia 21.
No. Jose, Pedro's cousin, was a witness to the crime, since he saw it take place. Jose was not suspected, even though he was in Leme on the 21st.

The addition of linguistic information to the specifications of the rhetorical relations and the application of our structure-sensitive syntactic rules clearly result in a significant improvement in text quality. Although this can be seen in the English translations, the extent of the improvement in our examples is much more obvious to readers unfamiliar with Portuguese.

Of course, a number of other, non-structural, considerations will also need to be taken into account for further improving the text style. For example, there is the general problem of reference. The improved version of Example 2 would be much better, in the sense of being more easily understood, if "in Leme on the 21st" were replaced with "present at the crime". Just when replacements of this type are required and what they should look like remains to be solved. Similarly, there is the ɐm of synonymy. It is a basic rule of good text that the same word or marked ʎntactic construction must not be repeated too often or too closely together. This is extremely important for Portuguese and, to a somewhat lesser extent, for English. Although LETTERA deals with this problem, it does not do so in a theoretically motivated way. General solutions must necessarily involve theoretically inspired rules for determining just what "too often" and "too close" really mean.

The generality of our rules to the full set of RST relations is not yet known since they have only been tested with the subset of relations used in LETTERA. Our examination of other relations leads us to believe that they have a wider application, but this issue will only be resolved by an in-depth study of all relations. This is the subject of ongoing research.

Finally, this work raises a theoretical point with respect to RST, in that applying RST as an analysis technique to the resulting text will not produce a

structure that is identical to that produced by the planner. The question arises as to what this phenomenon really means. The answer to this revolves around the issue of what is the status of the planner's RST structure. We would want to claim here that the structure produced by the text planner reflects the relations that hold between the information elements of the text to be produced and that although this must be related to the discourse structure of the text, it is not equivalent to it. The crucial test of the compatibility of the two structures rests on whether the second retains the hierarchical relations of primary and secondary information of the first. In this sense the planner's structure can be considered as the mental model we want the reader to have of the text content, a model which can be expressed in a number of ways.

References

[1] Grice, P. (1975). 'Logic and Conversation'. In P. Cole and J. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*, Academic Press.

[2] Mann, W.C. and Thompson, S.A. (1986). 'Rhetorical Structure Theory: Description and Construction of Text Structures'. Technical Report ISI/RS-86-174, Information Sciences Institute, University of Southern California.

[3] Mann, W.C. and Thompson, S.A. (1987). 'Antithesis: A Study in Clause Combining and Discourse Structure'. In R. Steele and T. Threadgold (Eds.), *Language Topics: Essays in Honour of Michael Halliday*, Vol. 2, John Benjamins Publishing Company, Philadelphia.

[4] Nunes, M.G.V. (1988). 'Deep Generation in a Crime Knowledge-Based System'. Series Monografias em Ciencias da Computacao, Dept. de Informatica, PUC/RJ.

[5] Nunes, M.G.V. (1989). 'LETTERA: um Gerador de Texto em Portugues'. Series Monografias em Ciencias da Computacao, Dept. de Informatica, PUC/RJ.

[6] Nunes, M.G.V. and Scott, D.R. (1988). 'O Componente Estrategico de um gerador de respostas para um sistema baseado em conhecimento criminal', Anais do 5 Simposio Brasileiro de Inteligencia Artificial, Natal.

[7] Hovy, E. (1988). 'Planning Coherent Multisentential Text'. *Proceedings of the 26th Meeting of the ACL*. Buffalo, New York.

[8] Matthiessen, C. and Thompson, S.A. (1987). 'The Structure of Discourse and "Subordination" '. In J. Haltman and S.A. Thompson (Eds.), *Clause Combining in Discourse and Grammar*, John Benjamins Publishing Company, Amsterdam.

[9] Lyons, J. (1968). *Introduction to Theoretical Linguistics*, Cambridge University Press, Cambridge.

[10] Frazier, L., Taft, L., Roeper, T., Clifton, C., Ehrlich, K. (1984). 'Parallel Structure: a Source of Facilitation in Sentence Comprehension', *Memory and Cognition*, 12.

[11] Scott, D.R. and de Souza, C.S. (1989). 'Conciliatory Planning for Extended Descriptive Texts', Series Monografias em Ciencias da Computacao, Dept. de Informatica, PUC/RJ.