Summarizing RST trees focusing on referential chains: A case study

Eloize Rossi Marques Seno[1, 2], Lucia Helena Machado Rino[1]

[1]Núcleo Interinstitucional de Lingüística Computacional – NILC/DC-USFCar
Rodovia Washington Luiz, km 235 – 13565-905 – São Carlos – SP – Brasil
http://www.nilc.icmc.usp.br

[2]Centro Universitário Central Paulista – UNICEP
Rua Miguel Petroni, 5111 – 13563-470 – São Carlos – SP – Brasil

{eloize@mail.fpte.br, lucia@dc.ufscar.br}

Abstract. *In this paper we introduce RHeSumaRST, a heuristics based system that aims at pruning an RST tree in order to yield its corresponding summary. Those heuristics focus especially on identifying when a discursive segment that embed an anaphoric term is chosen to compose the summary, in which case the discursive segment that embed its antecedent must also be chosen. To both address referentiality and rhetorical structuring, Veins Theory is added to RST Theory in order to drive pruning. A preliminary case study is presented that assesses pruning of RST trees resulting from discourse-analyzing texts written in Brazilian Portuguese.*

1. Introduction

In this article, we introduce RHeSumaRST (Heuristic Rules for Summarizing RST trees), a heuristics driven automatic summarizer that aims at pruning an RST tree in order to yield its corresponding summary (Seno and Rino, 2005). The RST tree of a text mirrors its discourse structure by means of RST relations between discourse segments. If a text is coherent, its RST tree consistently relates its discourse segments in such a way that the underlying message may be retrieved. Focusing on Automatic Summarization (AS), rhetorical relations may signal discourse segments that are potentially superfluous for exclusion through their satellites. This has already been suggested by others (e.g., Sparck-Jones, 1993; O'Donnell, 1997). However, pruning may not be blind, since excluding all satellites from an RST tree may introduce coherence breaks, besides implying significant loss of information.

The former issue is the focus of this work: to guarantee coherence, some satellites may be identified as essential for the summary. This applies, for example, to the phenomenon of referentiality: a coherence loss introduced by co-reference breaks is one of the most serious problems in AS. When a discourse segment that embeds an anaphor is chosen to compose a summary and its antecedent is not (henceforth, causing a dangling anaphor), a coherence break occurs, unless it is a direct anaphor[1]. Defining pruning heuristics based solely on RST relations does not prevent such a phenomenon

[1] Direct anaphors are those whose anaphoric term just mirrors the expression of its antecedent and, thus, poses a repetition.

to occur, since the RST Theory does not address it explicitly. Our RHeSumaRST system adds to RST the Veins Theory (Cristea et al., 1998), aiming at overcoming that.

The Veins Theory was chosen for its delimiting the domain of referential accessibility of a discourse unit, expressed by its vein. Veins are defined on RST structures to signal the scope of the discourse in which anaphora antecedents may occur. In this way, co-referent discourse segments that are included in the same vein may lead to a coherent discourse. Thus, the main premise in defining RHeSumaRST heuristics follows: pruning RST trees is based on identifying and excluding superfluous satellites, as usual, but only after assuring that discourse segments that are candidate to exclusion do not spoil the vein in which they are inserted. This, in turn, aims at keeping the summary RST tree coherent.

The only phenomenon of referentiality dealt with by RHeSumaRST is the co-referential chaining, i.e., the occurrence of both anaphoric term and its antecedent in the text. Only definite anaphors (Vieira et al., 2002) were considered, which are the ones signaled by nouns phrases. In Brazilian Portuguese, they are generally introduced by a definite article (e.g., ‘o menino’, or *the boy*). Henceforth, co-referential chains will be referred to by the acronym CRCs.

Figure 1 presents RHeSumaRST pipelined architecture: first, an input RST tree is annotated by applying Cristea et al.’s algorithm (1998) of delimiting veins; then pruning takes place[2]. Several heuristics may be applied to prune a source RST tree. Although linguistic realization is of utmost importance for RHeSumaRST, as it is the discourse analysis of real texts, both modules have not been addressed in our AS fundamental approach so far.

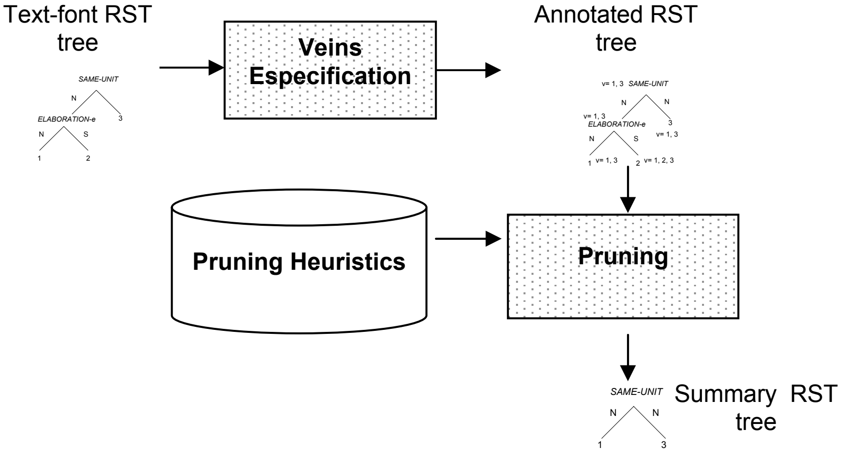

Figure 1: RHeSumaRST architecture

In Sections 2 and 3 we briefly outline the main features of RST and the Veins Theory, respectively. Then, we describe how the heuristics have been defined (Section

[2] Both modules have been implemented in collaboration with Leandro M. Hanada.

4), presenting a case study that assess them in Section 5, by addressing RHeSumaRST informativity and coherence. Final remarks are presented in Section 6.

2. The Rhetorical Structure Theory

According to the RST Theory (Mann and Thompson, 1987), an RST tree is composed by *Elementary Discourse Units* (EDUs) inter-related through rhetorical relations. These may be compositional: an RST relation may hold between two EDUs, resulting in an RST subtree. This, in turn, may also be related to another RST subtree. In this way, a coherent text will have no dangling RST subtrees. An RST tree is composed, thus, of internal nodes that are all RST relations and leaves as EDUs.

Significance is also addressed in RST: RST nuclei (Ns) are more significant, or relevant, than their satellites (Ss), and are introduced by mononuclear rhetorical relations. Equally significant discourse segments are multinuclear. For Text 1 (Figure 2, with EDUs numbered for reference), extracted from TeMário (Pardo and Rino, 2003)[3], its underlying RST tree is that presented in Figure 3[4] (some definite anaphors in bold).

[1] **A empresa Produtos Pirata Indústria e Comércio Ltda.,** de Contagem [2] (na região metropolitana de Belo Horizonte) [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente. [4] Os ganhos são atribuídos pela diretoria d**a fábrica** à nova filosofia que vem sendo implantada n**a empresa** desde outubro do ano passado, [5] quando **a Pirata** se iniciou n**o Programa Sebrae de Qualidade Total**.

[6] Dona de 65% do mercado mineiro de temperos, condimentos e molhos, **a Pirata** reúne atualmente 220 funcionários. [7] A coordenadora d**o programa de qualidade na empresa**, Márcia Cristina de Oliveira Neto, disse que [8] ainda não é possível dimensionar os ganhos financeiros que "certamente" **a empresa** terá, em conseqüência da melhoria da qualidade de seus produtos e serviços. [9] Por enquanto, os benefícios mais visíveis, segundo ela, são a organização e a limpeza d**a fábrica**. [10] "Também a relação entre as pessoas tem melhorado bastante. As informações estão mais claras e os funcionários e clientes, mais satisfeitos”.

Figure 2: Text 1

[3] Available in: http://www.linguateca.pt/Repositorio/TeMario (last access: february/2005).

[4] Actually, the current set of RST relations embedded in RHeSumaRST (c.a. 35) mixes original ones with some of those proposed by Carlson and Marcu (2001).

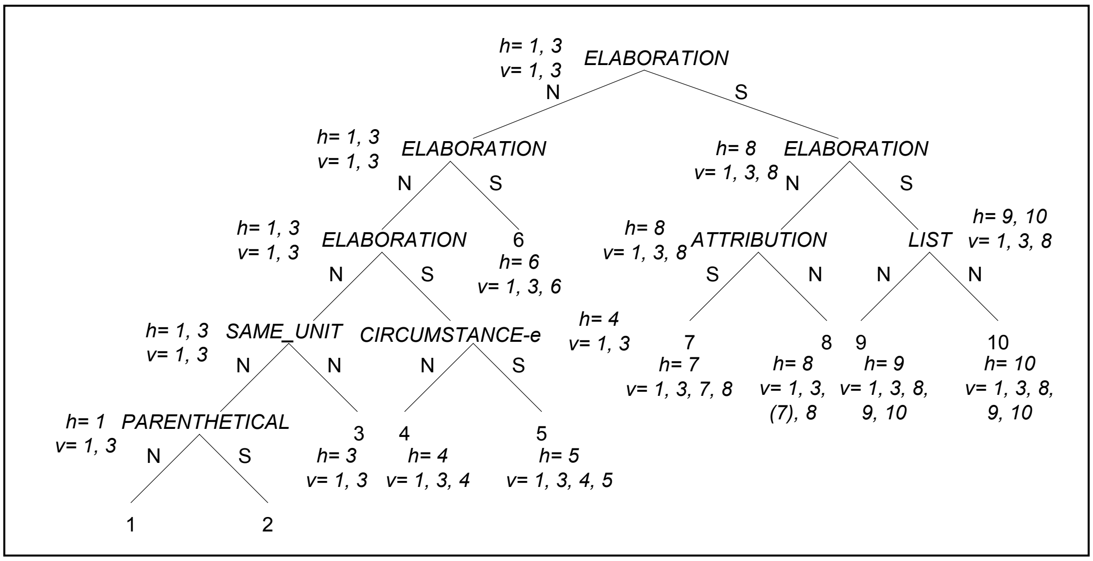

Figure 3: RST tree of Text 1

3. The Veins Theory

Based on RST nuclearity, the Veins Theory proposes to delimit the domains of referential accessibility of each EDU in an RST tree, resulting in the so-called “veins”. The vein of an EDU is defined as the set of discourse units that embed the antecedent of an anaphora related to that EDU. This definition gives rise to our main premise, as formerly pinpointed, which addresses co-referential chaining by verifying if a complete CRC is embedded in a unique vein.

Applying Cristea et al.’s algorithm to compute the veins of the RST tree of Text 1 yields the annotations also included in Figure 3 – *heads* (*h*) and *veins* (*v*) are presented in italics. A head of an RST node N is the set of its most salient EDUs in the discourse segment which embeds N; its vein is drawn on its head basis. For example, for node 1, h = 1 e v = 1, 3; for the RST subtree headed by SAME_UNIT, h = v = 1,3.

To illustrate the determination of the domain of referential accessibility, consider the EDU [5] in Text 1. According to Figure 3, its vein is composed of EDUs [1], [3], [4], and [5]. So, the antecedent of the anaphor *Pirata* occurring in [5] ought to be included in any EDU in the segment 1,3,4. Actually, it is embedded in [1] (the second reference anaphoric being in [4]), as reproduced below:

> [1] **A empresa Produtos Pirata Indústria e Comércio Ltda.**, de Contagem [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente.
>
> *(The industry Produtos Pirata Indústria e Comércio Ltda., from Contagem, will register this year an increase in productivity in its commercial and industrial areas of 11% and 17%, respectively.)*
>
> [4]-[5] Os ganhos são atribuídos pela diretoria da **fábrica** à nova filosofia que vem sendo implantada na **empresa** desde outubro do ano passado, [5] quando a **Pirata** se iniciou no Programa Sebrae de Qualidade Total.

(The gains are due by the board of the industry to the new philosophy that has being adopted in the industry since October last year, when the Pirata was introduced in the Sebrae Program of Total Quality.)

A possible summary considering the above could signal [5] as the cause of the industry progress and, so, as a salient information to preserve in the summary. As a result, the corresponding realization could be, for example:

[1] A empresa Produtos Pirata Indústria e Comércio Ltda. [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente, avanço significativo a partir da adoção do [5] Programa Sebrae de Qualidade Total.

(The industry Produtos Pirata Indústria e Comércio Ltda. will register this year an increase in productivity in its commercial and industrial areas of 11% and 17%, respectively, a significant progress since de adoption of the Sebrae Program of Total Quality.)

From the writer's viewpoint, this is a good summary, since it signals the industry progress ([1]-[3]), which is the main topic of the source text, and its cause ([5]). [4] was not included because, in our reading, [4]-[5] together may lead to the cause relationship and the phrase '*a significant progress since de adoption of*" overlaps [4]. So, whilst the analysis of the source text (Figure 3) indicates [5] alone as a circumstance of [4], that phrase introduces it as a cause of [1]. This is clearly licensed in our summarizing example, as a result of interpreting cause as another possible choice for [5] inclusion in the source text. In its illustrated RST tree, [5] is a satellite of the rhetorical relation CIRCUMSTANCE-e whose vein is signaled by EDUs v= 1,3,4,5. In building a summary RST tree including only [5], a very extreme summary could lead to 'A **Pirata** se iniciou no Programa Sebrae de Qualidade Total.' (*Pirata was introduced in the Sebrae Program of Total Quality.*), which conveys a complete message, but does not preserve the main idea of the source text. So, this would bring about a loss of both informativity and a topical move. For this reason, [1] must be included and, thus, also [3] (since this composes the same unit as [1]). Moreover, its meaning would only be adequately grasped if the reader knew what the noun *Pirata* meant. So, in order to keep the summary closer to its source with respect to both, informativity and coherence, a more adequate summary would have to include at least those EDUs embedded in the vein of EDU [5].

In other words, by reading [5] as the cause for the industry progress and, thus, a very important information, an automatic summarizer could just select the very same EDU to build up a one-sentence summary. This, however, would be significantly less informative, for its missing the main idea of the source text, which emphasizes its progress due to a quality program. So, it does not apply to introduce the industry program quality. A heuristics based upon including the antecedent of an anaphor would prevent this message to be conveyed, keeping it compatible with the source.

3. Heuristics based on co-references for AS of RST trees

The above example illustrates well the application of pruning heuristics in the RHeSumaRST system: they address both coherence and informativity by focusing on constraints that aim at not having coherence breaks introduced by particular choices of EDUs. Since RHeSumaRST does not resolve anaphors, the heuristics are driven towards including a complete vein, once one of its components is chosen.

Defining the set of pruning heuristics has been corpus-driven: the corpus was composed of 30 newspaper articles from TeMário and its analysis aimed at (a) identifying those RST satellites that were indeed superfluous; (b) verifying the contexts of co-referentiation that could introduce coherence problems in summarizing. The texts were pre-processed in three distinct phases, as follows: firstly, their RST trees were built with the *RST Annotation Tool*[5]; secondly, the veins of the resulting RST trees were automatically obtained; finally, the occurring CRCs were also annotated with the MMAX tool (Müller and Strube, 2001).

Aiming at (a) above, we compared each RST tree with the corresponding manual summary[6]: we verified if each EDU in an RST tree had corresponding information units. The underlying hypothesis here was that, by defining heuristics based on information common to the manual summaries, the heuristics would be able to recognize content judged relevant in the source text under summarization. The comparison aimed, thus, at guaranteeing minimum informativity in the automatic summaries. This methodology implies that heuristics be based on those RST relations that signal more significantly the content of interest, for any source text (it is not the aim of this paper to discuss genre dependence).

Our analysis showed that most mononuclear RST relations (c.a. 97%) had their satellites included in the manual summaries in 50% or less of the cases. Many of them had no satellite preserved at all, such as the CIRCUMSTANCE relation. Only EXPLANATION ARGUMENTATIVE had more than 50% of its satellites present in the summaries (57%). However, this RST relation is meaningless in the corpus (only 0.4% occurrences). These results may indicate that satellites of RST trees are indeed non-relevant for AS and, thus, should be directly excluded, in pruning mononuclear RST relations. Multinuclear ones also appear in the corpus. However, they were not our focus, because if we decide to include in a summary RST tree one of the EDUs of those relations, all of them should be included. So, there are no pruning heuristics for them.

Concerning goal (b), the corpus analysis aimed at identifying the domain of referential accessibility of definite anaphors occurring in the source text, in order to verify its structural correspondence with its RST tree and, thus, derive proper heuristics to guarantee that the summary will not have dangling anaphors. Then, we looked for both, its anaphoric and antecedent terms in its RST tree, to see if they were present in the same vein. The hypothesis here was that, if a complete chain were embedded in the same vein, heuristics should be based on the preservation of the full vein to guarantee the minimum of coherence of the summaries, concerning CRCs .

The results showed that, for 80% of the CRCs in the corpus, both anaphor and antecedent occurred in the same vein. For the corresponding RST relations, heuristics were thus defined that were limited to excluding only those satellites that were not in the domain of referential accessibility of the EDUs already chosen to compose a summary.

[5] Available in: http://www.isi.edu/~marcu/discourse/AnnotationSoftware.html (last access: march/2005).

[6] TeMário texts already come along their manual summaries, built by a professional writer. So, for evaluation purposes, they are our ideal summaries (Mani, 2001).

As a result of the corpus analysis, 30 pruning heuristics were defined, which compose the main module of the RHeSumaRST system, as described in (Seno and Rino, 2005). To evaluate them, we automatically produced summaries for 10 texts also selected from TeMário, having as input their source RST trees, as described in the next section. Although RHeSumaRST does not embed a proper linguistic realizer (see Figure 1), the summaries were obtained by just juxtaposing the leaves of its summary RST trees.

4. Assessing the heuristics: a case study

We assessed the heuristics aiming at verifying both, if they could identify the most relevant information of the source RST trees and if they could guarantee coherence in the summaries, concerning especially the occurrence of full CRCs. The former goal implies calculating the *degree of informativity* of the summaries; the latter, their *degree of coherence loss*. We carried out a similar pre-processing of the test corpus as that on the analysis of a similar corpus for defining the heuristics.

4.1. Evaluation on informativity

ROUGE tool[7] (Lin, 2004) was used in this phase, which allows to automatically calculating the degrees of informativity of the automatic summaries. The comparison is made between these and manual ones, also considered ideal. Measures to grade informativity are based on the co-occurrence of content units between both, ideal and automatic summaries, thus, on recall: the more an automatic summary recalls content of the ideal one, the more informative it is (number of content units common to both, automatic and ideal summaries divided by the total number of content units of the ideal one). Co-occurrence of content units may be based on diverse linguistic patterns, after customizing ROUGE: we used unigrams (hereafter, ROUGE-1), bigrams (ROUGE-2), and longer subsequences of usual words (ROUGE-L). The number of components in any n-gram in those sequences is fixed in ROUGE.

Ideal summaries were produced by five native speakers of Brazilian Portuguese, under a 70% compression rate, amounting to 50 summaries. So, each text of the test corpus had five ideal summaries to be compared with (in DUC'2004, for example, 4 ideal summaries were used). Automatic summaries were also produced under the same compression rate. Finally, recall was calculated between each automatic summary and all its five ideal ones by using ROUGE-1, ROUGE-2 and ROUGE-L.

We additionally compared RHeSumaRST performance with two other systems: that proposed by Marcu (1997, 2000), hereafter *Salience Model*, which selects only the most salient EDUs of an RST tree to compose a summary, and that which prunes every satellite of an RST tree, leaving only its nuclei, hereafter *Topline Model*. This has been named so because we consider that pruning all the satellites and leaving all the nuclei of a source RST tree (thus, only central information, according to Mann and Thompson) is very likely to provide a highly informative summary. This is confirmed in our test, as shown in Table 1 (average numbers given).

[7] A broad-coverage tool to evaluate summaries that was used in the last two DUC conferences (see www-nlpr.nist.gov/projects/duc/index.html (last access: march/2005)).

Table 1: Degrees of informativity of RHeSumaRST summaries with 5 ideal summaries

AS systems	ROUGE-1	ROUGE-2	ROUGE-L
Topline	0.58424	0.33659	0.55663
RHeSumaRST	0.57110	0.32640	0.54550
Salience	0.55757	0.32286	0.53192

Amongst the three systems, RHeSumaRST was the closest to Topline when ROUGE-1 and ROUGE-L were used. However, in using ROUGE-2, both RHeSumaRST and Salience had similar performances. These results show that, even excluding nuclear EDUs from the source RST trees, RHeSumaRST may be as informative as Topline, i.e., RHeSumaRST can depict better than Topline those content units that do not contribute to informativity.

We also performed a RHeSumaRST assessment using only 3 ideal summaries. However, this decrease implied a considerable decrease of its average recall, in spite of its keeping outperforming the Salience system in ROUGE-1 and ROUGE-L, as shown in Table 2.

Table 2: Degrees of informativity of RHeSumaRST summaries with 3 ideal summaries

AS systems	ROUGE-1	ROUGE-2	ROUGE-L
Topline	0.56431	0.32326	0.53402
RHeSumaRST	0.53738	0.29929	0.51080
Salience	0.52877	0.30084	0.50245

4.2. Assessment of coherence

In this assessment, each automatic summary was manually compared with its corresponding source text, in its annotated version. The goal was to verify if there was no dangling anaphor, thus, no coherence break in the summary. Once identified a coherence break in the automatic summary, we searched for the corresponding text span in the source text to verify if its corresponding content unit was embedded in a CRC. In this case, we retrieved its antecedent and went back to the summary, to check if it were included. In other words, we certified that a dangling anaphor was the cause of the coherence break. No coherence break was added for direct definite anaphors because they do not introduce dangling anaphors.

The same systems used in the previous evaluation were also used here. Table 3 shows the number of coherence breaks of each and its representativeness in the corpus.

Table 3: Coherence breaks in RHeSumaRST summaries

AS systems	# of CRCs	# of coherence breaks	coherence breaks (%)
RHeSumaRST	93	5	5
Topline	89	7	8
Salience	81	12	15

In a way, having more coherence breaks in the Topline and Salience systems is understandable, since they do not address explicitly the means to include information in summaries that may contribute to coherence. Particularly, Salience does not embed any prevailing resource to preserve an antecedent of a chosen anaphor. This may justify its worst performance. Topline rate may just indicate that the RST trees mirror well the organization of the source texts, i.e., that they are consistently related with respect to RST nuclearity, hence their coherence may be assured almost independently of their satellites.

If, on one hand, RHeSumaRST presented the least rate of coherence breaks, signaling its usefulness to deal with coherence problems, on the other hand the results are too close to the others to draw a meaningful conclusion of the reported experiment. Moreover, our test corpus is very small. Limiting it is understandable, for the huge effort of building source RST trees by hand and carrying out a human evaluation. However, RHeSumaRST will be useful for real AS in the near future, after plugging to it DiZer, a discourse analyzer of texts written in Brazilian Portuguese (Pardo et al., 2004).

5. Final remarks

Although RHeSumaRST demanded a very effortful work on defining its heuristics, it adds to fundamental approaches the advantages of both, the RST nuclearity and the Veins Theory domains of referential accessibility. Clearly, for certain genres whose texts do not embed a significant amount of co-referential chains, the proposed model would be too sophisticated. However, in assuring that a vein will be completely reproduced in the summary RST tree, RHeSumaRST may also be used independently of that phenomenon. Actually, it does not recognize automatically the occurrence of an anaphora. This is the reason for adding the Veins Theory, which is its main contribution to fundamental AS.

It is also noticeable that, for most genres and languages, including Brazilian Portuguese, co-referential chaining is one of the most prominent ways of enriching texts. This justifies RHeSumaRST usefulness. Its above coherence performance for 10 texts may well be due to inadequate RST structuring or veins annotation. Actually, the 5% value obtained in Table 3 is due to the lack of a local inter-relationship between the corresponding anaphoric and antecedent EDUs in their source text: they do not occur in the same vein. In turn, this is due to the absence of a direct structural relationship between both, which is not addressed by the Veins Theory.

References

Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual.* Technical Report ISI-TR-545, University of Southern California.

Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory: A Model of Global Discourse Cohesion and Coherence. *In the Proceedings of the Coling/ACL' 1998*, pp.281-285. Montreal, Canadá.

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *In the Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, Barcelona, Spain.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD Thesis, Department of Computer Science, University of Toronto.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Müller C. and Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. *In the Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, pp. 90-95.

O'Donnell, M. (1997). Variable-Length On-Line Document Generation. *In the Proceedings of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duiburg, Germany.

Pardo, T.A.S. e Rino, L.H.M. (2003). *TeMário: Um Corpus para a Sumarização Automática de Textos*. Série de Relatórios Técnicos: NILC-TR-03-09, ICMC/USP, São Carlos-SP.

Pardo, T.A.S.; Nunes, M.G.V.; Rino, L.H.M. (2004). *DiZer: an Automatic Discourse Analyser for Brazilian Portuguese*. In the *Proceedings of the XVII Brazilian Symposium on Artificial Intelligence* - SBIA2004. São Luís, Maranhão, Brazil.

Seno, E.R.M. e Rino, L.H.M. (2005). *Heurísticas de Sumarização de Estruturas RST*. Série de Relatórios Técnicos: NILC-TR-05-04, ICMC/USP, São Carlos-SP.

Sparck Jones, K. (1993). *Discourse Modelling for Automatic Summarising*. Tech. Rep. No. 290. University of Cambridge, February.

Vieira, R.; Salmon-Alt, S.; Schang, E. (2002). Multilingual Corpora Annotation for Processing Definite Descriptions. *In the Proceedings of the Portugal for Natural Language Processing – PorTAL – 2002*, Faro, Portugal.