# Insights for better RST segmentation of texts in Portuguese?

**Lucia Helena Machado Rino[1], Ariani Di Felippo[2], Thiago A. Salgueiro Pardo[3]**

Núcleo Interinstitucional de Linguística Computacional – NILC

[1]Departamento de Computação, Universidade Federal de São Carlos

[2]Departamento de Letras, Universidade Federal de São Carlos

[3]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

lucia@dc.ufscar.br, arianidf@gmail.com, taspardo@icmc.usp.br

***Abstract.*** *Manually annotating a corpus of news texts in Portuguese – the CSTnews Corpus – led to questioning RST discourse segmentation and subsequent ways of handling RST structures for automatic summarization. In this article both aspects are considered under just one phenomenon: the occurrence of relative clauses in sentences that may be chosen to compose the automatic summaries. Specific guidelines for RST discourse segmentation have been applied to CSTnews. Here we show evidences that the distinction between explicative and restrictive relative clauses, as proposed by the Portuguese grammar, should be pursued for summarization modeling.*

***Resumo.*** *A anotação manual de um corpus de textos jornalísticos em português – o Corpus CSTnews – levou-nos a questionar a segmentação de textos baseada na Teoria RST e as possibilidades de lidar com estruturas RST para a sumarização automática. Neste artigo ambos os aspectos são considerados sob um só fenômeno: a ocorrência de orações relativas em sentenças que podem ser escolhidas para compor os sumários automáticos. O manual de segmentação RST foi usado para o CSTnews. Aqui relatamos as evidências de que a distinção entre as orações relativas explicativas e restritivas, segundo a gramática da língua portuguesa, deveria ser considerada para a modelagem de sistemas de sumarização.*

## 1. Introduction

For segmenting and analyzing texts based on RST [Mann and Thompson, 1987], NILC research group on Automatic Summarization (AS) adopted Carlson and Marcu's (2001) guidelines under the SUCINTO[1] and SUSTENTO[2] Projects. Regardless the diversity of proposals for segmenting texts[3], most researchers agree that non-overlapping text segments can be characterized as elementary discourse units (EDUs). These are usually conveyed by clauses, hence intra-sentential segmentation takes place, which is adequate for AS and Muti-document AS (MAS): it allows both for dealing with information in a more detailed way and assuring a low rate of information loss, two important AS conditions. In fact, if sentences are taken as minimum units, selecting and condensing text segments for MAS becomes more limited, in that it may either prevent, e.g., omitting details from a sentence, or endanger compression rate. As a result, sentence

[1] http://www.icmc.usp.br/~taspardo/sucinto/

[2] http://www.nilc.icmc.usp.br/~arianidf/sustento/

[3] See [Carlson et al., 2001; Nicholas, 1994] for a good account on this.



*Anais do IV Workshop "A RST e os Estudos do Texto"*, páginas 30–39, Fortaleza, CE, Brasil, Outubro 21–23, 2013. 

granularity may risk summary quality. For MAS redundancy or fusion treatment, for example, looking for lexical repetitions or clause removal are interesting approaches that may not be properly handled at the sentence level. In spite of that, sentence granularity is usually adopted due to its simplicity in dealing with text, mainly when superficial and extractive methods are considered (see, e.g., Jorge, 2010). However, when models for AS or language representation need deeper discourse organization, a clause-based granularity should be considered [Mann and Thompson, 1987; Carlson and Marcu, 2001; Taboada and Mann, 2006]: the semantic content of clauses would better correspond to EDUs, for retrieving the discourse structure underlying a text.

In this paper, we focus on the *restrictive* and *explicative* sub-types of relative clauses, which depict distinct functions in discourse. Hereafter they are correspondingly referred to as RESTRs and EXPLICs. Even when the same set of words applies to a relative clause, distinguishing them may imply switching its semantics. Texts 1 and 2 exemplify this[4]. Both enclose the same realization for relative clauses, but Text 2 adds to Text 1 just a comma preceding the relative pronoun 'que'. According to the traditional grammar, that comma signals the beginning of an *explicative* clause. Text 1, in turn, contains a *restrictive* clause. An experienced reader may realize that interpreting both texts either in Portuguese or in English certainly yields distinct messages concerning the referenced man who was passing by: Text 1 suggests that there might be several men, but just one is a participant of the reported event; Text 2 suggests that there is only one man; the embedded clause just adds more detail about him.

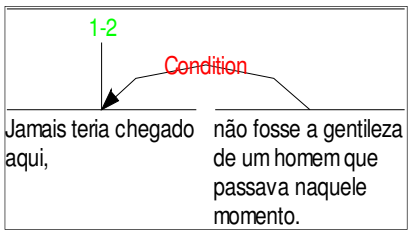
**Text 1.** Jamais teria chegado aqui, não fosse a gentileza de um homem [**que passava naquele momento**].
(*Never one would have arrived here, if it weren't for the kindness of a man [**who was passing by at that time**].*)

**Text 2.** Jamais teria chegado aqui, não fosse a gentileza de um homem, [**que passava naquele momento**].
(*Never one would have arrived here, if it weren't for the kindness of a man, [**who was passing by at that time**].*)

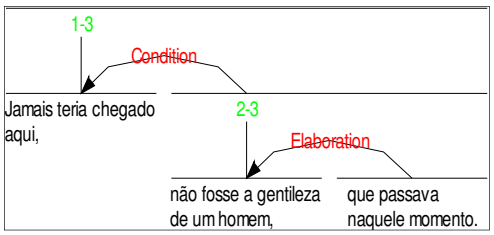
RST analysis by Mann and Thompson (1987) agrees with the above semantic variations, concerning the referred 'homem' entity. Thus, diverse RST trees would result from those relative clauses: RSTtree 1 shows that the RESTR clause in Text 1 comes along its main clause, just as one satellite of the CONDITION relation. RSTtree 2 embeds the EXPLIC clause as a satellite of the ELABORATION relation instead. Notice that Carlson and Marcu's segmentation protocol would yield just RSTtree 2 for both texts, ignoring that Text 1 refers to a specific description of 'homem'. Even considering the unfolded ELABORATION-ADDITIONAL (for EXPLICs) and ELABORATION-OBJECT-ATTRIBUTE (for RESTRs), as proposed for the annotation of the RST Treebank, the problem remains: satellites of the corresponding RST structures would refer to those EDUs that might not be candidates for exclusion. In other words, over-specifying ELABORATION would not suffice to prevent restrictive EDUs to be omitted from final summaries. Additional problems would also appear in adopting that over-specification, once both specific elaborations do not apply exclusively to semantic content conveyed by relatives.

[4] Literal English versions follow, for understanding, and show that the phenomenon is equally treated in that language.

As pinpointed, ignoring that distinction may introduce severe damages for Natural Language Processing (NLP), in general, and for AS in special. One may clearly realize that taking RSTtree 2 as the basis for building a Text 1-like summary would provide a misleading message if the 3rd EDU were considered superfluous, as the ELABORATION relation may entitle.




**RSTtree 1. RST tree of Text 1**




**RSTtree 2. RST tree of Text**

Considering other NLP scenarios, e.g., Natural Language Generation, linguistically realizing RSTtree 2 for Text 1 through the same grammar rules used for interpreting could just yield Text 2, whose message is clearly different from that conveyed by Text 1. This makes evident that segmenting clauses should not be evenly tackled for both texts, because the messages conveyed are not the same[5]. More importantly, taking for granted that satellites of elaborations may be omitted from a summary without diminishing its readability or altering its content [Webber et al., 2012] does not apply to this case. Therefore, there seems to be plenty of evidence that ignoring the different discourse functions of RESTRs and EXPLICs is misleading.

Despite the above argumentation, our CSTNews RST treebank [Cardoso et al., 2011][6] is based on the "relative clause detaching" rule [Carlson and Marcu, 2001], i.e., every relative clause is segmented, being it restrictive or explicative. We justified such an option by stressing that, for some cases in Portuguese, it was difficult to distinguish RESTRs from EXPLICs.

The focus of this paper is on questioning the burden of ignoring that distinction. In this paper we show evidences that the distinction between explicative and restrictive relative clauses should be pursued for summarization modeling. We show that, contrarily to our initial assertion, it should not be difficult to distinguish those cases even automatically, when traditional grammar rules are considered and a small amount of cues are used for that. Our research question is thus the following:

When a subordinate restrictive clause modifies the meaning of a component of the previous, main clause,
should it be considered as an independent, embedded clause?

Our posed claim is that, ignoring both functions of relative clauses when segmenting texts, a faulty and non-reliable RST structure results. After briefly describing general approaches to relative clause segmentation (Section 2), that is further elaborated taking the AS context into consideration and targeting only the phenomenon occurring in texts written in Portuguese (Section 3). We pursued certifying our claim by analyzing the occurrences of relative clauses in the CSTNews Corpus, as described in Section 4, where we also show that the phenomenon occurs similarly to both English and

[5] See [Souza and Scott, 1990] for a deep discussion on proper surface realizations based on RST trees.
[6] Corpus available at http://www.icmc.usp.br/~taspardo/sucinto/cstnews.html

Portuguese. Evidences emerging from that may throw insights for more promising RST structuring and handling in AS or other NLP contexts, as pinpointed in Section 5.

## 2. Segmenting relative clauses

Segmenting relative clauses as proposed by Carlson and Marcu tackles the low-level structure of the discourse, especially concerning the semantic content of two particular units: the main and the relative ones. Many approaches for automatically delimiting text spans refer to breaking a text into discourse-relevant units based on lexical, syntactic, and semantic information. They aim at discourse structuring mostly addressing high-level structures or assuming sentences or clauses as relevant fine-grained units, but usually they do not account for functions such as those introduced by RESTRs or EXPLICs (see Webber et al., 2012, for a review on that). Examples include tokenizing texts into EDUs [Carlson et al., 2001] or tokenizing discourse after sentence-level parsing [Polanyi et al., 2004]. The principles for building the RST Treebank also enroll relative clauses just as the embedded ones, which are usually recognized by containing a verbal element and being introduced by a relative pronoun or a quantifier that acts as that pronoun, e.g., *some of which*, *a number of which* [Carlson et al., 2001; Carlson and Marcu, 2001]. There is no concern about the inadequacy of considering RESTRs as embedded clauses as well.

In both English and Portuguese, usual relative pronouns that signal the beginning of an adjective clause are *who* (que, quem)*, which* (que)*,* and *that* (que). Other lexical items, or even other types of phrasing, may signal relative clauses as well, as we report in Section 4. In Portuguese there is no semantic distinction between *who, which* and *that,* as there is in English: all may be realized by the same pronoun 'que', which may introduce either RESTRs or EXPLICs. Determining function is based on the following rule: if omitting a relative clause changes the basic meaning of the main clause, it shall not be delimited by commas. So, EXPLICs must be preceded by commas, and RESTRs must not [Decat, 2010]. More importantly, EXPLICs carry *more detail* about the referred noun appearing in the previous clause, whilst RESTRs *constrain that noun meaning*. This means that RESTRs are not *supplementary* like EXPLICs. Both English and Portuguese grammars reinforce that under the functional perspective: the only actual embedded clauses should be the *explicative, appositive* ones[7].

## 3. The problem: Segmenting relative clauses for RST structuring

Once EDUs were determined, RST tagging CSTNews implied linking together the adjacent spans via rhetorical relations, for producing the final RST structure for each text. In RST, a mononuclear relation indicates that its nucleus is more salient to the discourse structure than its satellite. It turns out that, for RST-based AS models, those satellites may be candidates for exclusion from the intended summaries [Sparck-Jones 1993; O'Donnell, 1997; Seno and Rino, 2005, etc.].

Our problem is, thus, that, in pursuing Carlson and Marcu's guidelines for text segmentation, every restrictive information is entitled for suppression from final summaries when AS is considered, because they appear as satellites of ELABORATION relations. Clearly, when the main EDU is chosen for inclusion and its corresponding RESTR EDU is not, this may deteriorate the summary: a more generic meaning for the

[7] We thank Dr. Maria Beatriz Decat for valuable insights on this issue.

referred entity than it has been originally intended will be conveyed. Barely this would be a wise procedure for AS or MAS, unless traditional or functional perspectives for Portuguese or English are revised.

In arguing in favor of considering a more fine-grained typology of relative clauses and limiting our research topic only to segmentation aiming at building RST structures, we wondered how representative that phenomenon is in CSTNews. We also questioned how flexible our procedure should be, in the light of Carlson et al.'s concerns on consistency (2001, p. 3). That is, should their argument in favor of a "sacrifice" for delimiting phrases be adopted in our CSTNews annotation? Other researchers also claimed that subtypes should not be "treated as separate relations because in many cases in text they are not distinguishable" (in the RST homepage). This applies to RST ELABORATION. However, in ignoring the constraints that more fine-grained units impose to discourse, how severe a deviation from grammar and functional views that might be, in preventing discourse organization to properly mirror those?

We argue that the only way to assure that texts will be properly RST-structured is by taking into account the relative clauses sub-specification. To certify this, and reiterate reported evidences that relative clauses ought to be differentiated [Scott and Souza, 1990], we first annotated CSTNews for relative clauses, our next reported topic.

## 4. Distinguishing relative clauses in CSTNews

We distinguished RESTRs from EXPLICs in CSTNews entirely based on the referred linguistic assumptions. Firstly, we looked for any surface delimiters. A few were found in CSTNews, as we will soon report here. Although delimiting the clauses could be automatic, analyzing the context of their occurrence would bring more light into their function diversity. Thus, we manually annotated all the documents with RESTR and EXPLIC tags. We also registered along each tag their signaling cues, to have an overall account on the lexicalizations present in the corpus. We realized that there had been no automatic semantic parser that could meet our requirements for this fine-grained analysis, despite the existence of some, e.g., [Bick, 2000; 2007].

Actually, this procedure is in line with discourse chunking [Webber et al., 2012], in that we just looked for contiguous clauses, and not for the full text, for setting the limits between the main clause and its adjective relative one[8]. This is also in line with Marcu and Echihabi's approach (2002), but that relies an automatic and unsupervised processing for both delimiting subordinate clauses in general and determining the RST relations between them. To replicate the first task for corpora in Portuguese we should have both expressive raw and parsed data, but that may be not worthy for elaborations: Marcu and Echihabi show that their classifiers can well distinguish relations that differ from those (actually, they confirm that ELABORATION relations are too ill-defined, as do Knott et al. 2001). Thus, relative clauses are not properly handled either. Actually, Taboada and Mann (2006) advocate in favor of adopting just generic elaborations, opposed to adopting sub-specifications of any type, as those defined by Marcu (2000) or as the generic:specific one referred to in the RST homepage[9]. Clearly, there has been much attention that, for generic purposes, adopting a fine-grained treatment of

[8] In the Penn Discourse TreeBank, they should respectively correspond to *Arg1* and *Arg2,* being the latter the argument that is syntactically bound to the connective.
[9] http://www.sfu.ca/rst/01intro/definitions.html (May 20, 2013)

ELABORATION relations is not recommended. However, we aimed at proving that ignoring those relative clauses specificities for AS purposes may be harmful for properly preserving original messages. Since approaches tackling the task of identifying arguments to discourse connectives (like, e.g., Webber et al., 2012), do not focus on data in Portuguese, manually annotating CSTNews was a previous and mandatory task for exploring further other modern language technologies.

### 4.1. Data description

A new corpus tagged for RESTRs and EXPLICs resulted from CSTNews, which comprises 50 clusters of news texts (140 in all). There are 446 cases of relative clauses, cued as shown in Table 1: only those preposing pronouns or cues were found in the corpus, yet not necessarily in all their forms (e.g., corresponding to *in which*, only 'em que' occurs). Those cues apply similarly to both Portuguese and English, although the specifity levels may vary. Portuguese differs from English in gender and number inflections of the same lexical item, as shown ('cujo'/'cuja'; 'os quais'/'as quais', etc.). Even *where* (onde), which is an adverb, in some cases act as pronouns, i.e., when they mean the same as 'em que'.

**Table 1. Portuguese cues signaling relative clauses**

| Portuguese cues | Corresponding English cues |
|---|---|
| que | *who/which/that* |
| onde (corresponding to 'em que') | *where* (corresponding to *in which place*) |
| cujo, cuja, cujos, cujas | *whose,whom* |
| o qual, a qual | *the one which* |
| os quais, as quais | *the ones which* |
| em que, naquele que, na qual, no qual, nas quais, nos quais | *in which* |
| de que, dos que | *of which* |

As mentioned, all typical pronouns that signal relative clauses in English (1st line of the table) amount just for 'que' in Portuguese. Thus, it is not possible to distinguish clause subtypes like in English (with *which* or *that*). 'que' is also used with varied functions in Portuguese, which makes the task of identifying when it signals a relative clause even more subtle. However, only the plural form of 'o qual' appeared in the corpus. *Who* and *where* are also pronouns that may address either EXPLICs or RESTRs in both languages. Adverbs or adverbial fragments can also be noun qualifiers, thus cueing *adverbial type relative clauses* through 'onde'. No relative pronoun can be elliptical in relative clauses in Portuguese, like in *This is the man I saw*. It should always be explicited: "Esse é o homem **que** eu vi." Apart from well-marked cases, there are some in the corpus that clearly refer to EXPLICs, but they are not preceded by commas. Those were tagged "noCOMMA". Text 3 (from the CSTNews D4_C27_JB.txt file) illustrates this: in a soccer context, there is just one ball, which certifies that the relative clause is explicative. Table 2 shows the representativeness of actual cases occurring in CSTNews.

> **Text 3.** A bola [que aparentemente iria para fora] mudou de direção e foi parar no fundo da rede.
> (*The ball [which apparently would go off] changed direction and ended inside the goal.)*

We can see that 'que' is the typical delimiter of relative clauses in our corpus. Only 'que' and 'onde' signal both RESTR and EXPLICs, and the remaining cues are non-significant in both cases. The percentage of occurrence of RESTR cases, on itself, should be highly indicative of the need to distinguish those from the EXPLICs, for their function to be properly addressed. However, distinguishing them, as already stressed, was neither

carried out in the RST Treebank, nor in CSTNews. To still certify that this should be revised for its relevance to discourse representation, we explored further other data supplied along with CSTNews, as we describe next.

**Table 2. CSTNews representativeness of delimiting cues for relative clauses**

| RESTR | | | | | EXPLIC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| que | onde | de que | em que | dos que | que | onde | os quais | cujo/a | nas/nos quais | noCOMMA |
| 212 | 5 | 2 | 6 | 1 | 191 | 13 | 1 | 9 | 2 | 4 |
| **TOTAL RESTRs: 226** | | | | | **TOTAL EXPLICs: 220** | | | | | |
| **51% of all relative clauses** | | | | | **49% of all relative clauses** | | | | | |

## 4.2. Correlating relative clauses in both CSTNews and human multi-document extracts

CSTNews also comprises extracts produced manually by simulating MAS for each cluster, e.g., the file "C1_extrato_humano" in the Summaries folder of the cluster C1 amounts for those relevant sentences extracted from any of the source texts included in C1 (each extract also has explicitly each sentence source). All of them were built based on the corresponding human abstracts that also come with the corpus. Humans selected as many full sentences from the source texts as those present in their abstracts. For this reason, distinguishing the types of relative clauses in the extracts may be misleading: including full sentences in the summaries does not imply that their embedded relative clauses would be also selected; they might be discarded otherwise. In spite of this, correlating relative clauses found in both raw source texts and their corresponding human extracts may be a good approach to estimate the actual data distribution. We also perform similar correlation on a sample of freely produced abstracts for the same corpus in order to produce more realistic figures, as explained latter.

In CSTNews, each document may present sentences with more than one relative clause, and this equally applies to the human multi-document extracts. These amount to 411 sentences. To investigate the relevance of distinguishing RESTR from EXPLICs, we searched for the 446 cases that embed relative clauses in those human extracts. This was only possible because the manual task simulated automatic extracting; thus, the minimum unit considered could be matched with the searched one. Sentence 1 (from the multi-document human extract for cluster C3), e.g., shows this: the 1st clause is RESTR; the 2nd is EXPLIC.

> **Sentence 1.** A falha no reversor -- mecanismo **que** ajuda o avião a frear -- foi detectada pelo sistema eletrônico de checagem da própria aeronave, **que** continuou voando nos dias seguintes, com o reversor direito desligado. <D1_C3_Folha; p3-s1>

Overall, we found 90 relative clauses in 77 segments that embed relative clauses in the human-produced set, as shown in Table 3: 22% of the extracts do not present any relative clause.

**Table 3. Coinciding RESTR and EXPLICs in human multi-document extracts**

| # relative clauses | Amount of RESTRs | Amount of EXPLICs |
|---|---|---|
| 90 | 50 | 40 |
| **100%** | **56%** | **44%** |

One may see that the differing occurrences of RESTR and EXPLICs in human extracts is not that significant, although RESTRs seem more preferable for humans: they occur nearly ¼ times more often than the EXPLICs. If the granularity level were not sentential,

one might certify the expectation for EXPLICs to be omitted from summaries, once they have a detailing function of a noun entity.

### 4.3. Correlating relative clauses in both CSTNews and human abstracts

To certify that actually humans would rather prefer to keep RESTRs instead of EXPLICs, we manually checked the occurrences of RESTRs and EXPLICs in the real abstracts that served as bases for those human-extracts. In doing so, we also should be able to detect possible distortions in the above results. After all, manually produced extracts replicate the well-known problems of extractive automatic summarization. Since it is not simple to match abstract sentences with their source counterparts (due to abstracting being a rewriting task), we checked only the occurrences of 'que', the most used cue for relative clauses in CSTNews. Table 4 summarizes the results.

**Table 4. RESTR and EXPLICs in human multi-document abstracts**

| # relative clauses | Amount of RESTRs | Amount of EXPLICs |
|---|---|---|
| 50 | 35 | 15 |
| **100%** | **70%** | **30%** |

From 50 occurrences of 'que' in the abstracts, 70% RESTRs against 30% EXPLICs demonstrates now that the much higher frequency of occurrence of RESTRs (they occur 133% times more often than the others) indeed makes evident that humans selected them to compose their abstracts, but put aside EXPLICs. Going after the other delimiters of relative clauses is most probably as non-significant as they are in the CSTNews source texts (see Table 2). Still, it remains to be seen if the reported representativeness would change when the full corpus is considered.

## 5. Subclassifying relative clauses: Are there enough insights for better RST segmentation of texts in Portuguese?

In all the cases reported above, RESTRs appear more often than EXPLICs in diverse corpora included in CSTNews. Through the first analysis, we could account for relative clauses occurring in all its 140 source texts that are grouped into 50 clusters. Then, to verify if the outperforming of RESTRs over EXPLICs were not accidental in MAS, we compared the sources with the human multi-document extracts, assuming that human choices would be based on the unfolded restrictive and explicative functions of relative clauses. However, having a sentence as the minimum unit in this corpus of human extracts could be a burden, because it becomes unfeasible to detect if EXPLICs appear in the extracts by choice, or else, if humans just replicate RESTRs as they appear in the corresponding source documents because they have no other choice. To overcome that, we proceeded to the 3rd analysis: in looking for relative clauses in human abstracts, we could indeed certify if such choices were purposeful, because, in their rewriting task, human abstractors could use their subjective judgments for sentence relevance and, occasionally, leave aside RESTRs, or include EXPLICs. In other words, they could perform abstracting the other way around. As a result, our AS models, and even our claim, would fall apart.

Whilst comparing real data with human multi-document extracts could be considered fragile, having humans freely selecting restrictive information significantly more often than selecting explicative ones to compose their abstracts legitimates the former correlation. Moreover, it throws even more reliability on our theoretical

assumptions and certainly corroborates the views of both English and Portuguese grammars. After all, we can certify that

Subordinate restrictive clauses *should not* be considered independent, embedded clauses.

Although the distinction discussed here may be crucial for AS or MAS, having RESTRs segmented as they are in the CSTNews Corpus may be too severe for most NLP applications, as we pinpointed earlier. Actually, the distribution of RST relations in the CSTNews corpus shows that elaborations occur significantly higher than any other relation. Certainly that number would decrease if RESTRs were properly addressed. Oppositely to that, if satellites of ELABORATIONs did not mirror RESTRs at all from source texts, their expressive inclusion in any type of summary should be even higher than that showed in the previous section. Thus, by pursuing insights from different sources, there is practical evidence that supports a more theoretically-based methodology for RST segmentation and the organization that follows, concerning the special sub-types of relative clauses. A follow-up step would thus naturally result: revising CSTNews RST segmentation and structuring, aiming at using it for training classifying models for MAS. It also remains to future work another, but no less interesting problem: dealing with reduced relative clauses, which have the same functionalities as the others reported in this paper.

## Acknowledgements

The authors are grateful to FAPESP for supporting this work.

## References

Bick, E. (2000). *The parsing system PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD Thesis. Arhus University, Arhus.

Bick, E. (2007). Automatic Semantic Role Annotation for Portuguese. In *Proc. of the 5th Workshop on Information and Human Language Technology, XXVII Congresso da SBC*, pp. 1713-1716. Rio de Janeiro, RJ.

Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In Proc. *of the 3rd RST Brazilian Meeting*, pp. 88-105. STIL 2011. October 26, Cuiabá/MT, Brazil.

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

Carlson, L.; Marcu, D.; Okurowski, M.E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proc. of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Aalborg, Denmark.

Decat, M.B.N. 2010. Relações retóricas e funções textual-discursivas na articulação de orações no português brasileiro em uso. *Calidoscópio,* Vol. 8, n. 3, pp. 167-173.

Jorge, M.L.C. (2010). *Sumarização Automática Multidocumento: Seleção de Conteúdo com base no Modelo CST (Cross-document Structure Theory).* MSc. Dissertation. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. April, São Carlos,SP, Brazil. 86p.

Jorge, M.L.C.; Pardo, T.A.S. (2010). Experiments with CST-based Multi-document Summarization. *In Proc. of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82. July 16, Uppsala, Sweden.

Knott, A.; Oberlander, J.; O'Donnell, M.; Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren (eds.), *Text representation: linguistic and psycholinguistic aspects*, pp. 181–196. Benjamins.

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

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

Marcu, D.; Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In the *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 368-375. Philadelphia.

Nicholas, N. (1994). *Problems in the Application of Rhetorical Structure Theory to Text Generation.* Unpublished Masters' Thesis. University of Melbourne, Melbourne.

O'Donnell, M. (1998). Variable-Length On-Line Document Generation. *In Proc. of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duiburg, Germany, 1997.

Polanyi, L.; Culy, C.; van den Berg, M.; Thione, G.L.; Ahn, D. (2004). Sentential Structure and Discourse Parsing. In *Proc. of the ACL 2004 Workshop on Discourse Annotation,* Barcelona, Spain (apud Webber et al., 2012).

Scott, D.R.; Souza, C.S. (1990). Getting the Message Across in RST-based Text Generation. In R. Dale, C. Mellish, and M. Zock (eds.), *Current Research in Natural Language Generation,* pp. 47–73. London, Academic Press.

Seno, E.R.M.; Rino, L.H.M. (2005). Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In H. Saggion (ed*.), Proc. of the Workshop on Crossing Barriers in Text Summarization Research Recent Advances in Natural Language Processing* (RANLP'2005), pp. 70-75. Borovets, Bulgaria.

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

Taboada, M.; Mann, W.C. (2006). Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies*, 8(3), pp. 423-459.

Webber, B.; Egg, M.; Kordoni, V. (2012). Discourse Structure and Language Technology. *Natural Language Engineering* 18(4): 437–490. October 2012. Cambridge University Press.