CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese

Paula C. F. Cardoso[1], Erick G. Maziero[1], Maria Lucía R. Castro Jorge[1], Eloize M. R. Seno[2], Ariani Di Felippo[3], Lucia H. M. Rino[4], Maria das Graças V. Nunes[1], Thiago A. S. Pardo[1]

Núcleo Interinstitucional de Linguística Computacional – NILC

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Avenida Trabalhador são-carlense, 400 - Centro
13566-590 - São Carlos/SP, Brazil

[2]Instituto Federal de São Paulo
Rodovia Washington Luís, Km 235, AT-6, Room 119
13565-905 - São Carlos/SP, Brazil

[3]Departamento de Letras, Universidade Federal de São Carlos
Rodovia Washington Luís, Km 235, P.O.Box 676
13.565-905 - São Carlos/SP, Brazil

[4]Departamento de Computação, Universidade Federal de São Carlos
Rodovia Washington Luís, Km 235, P.O.Box 676
13.565-905 - São Carlos/SP, Brazil

{paulastm,egmaziero,castrito21,eloizeseno}gmail.com, ariani@ufscar.br,
lucia@dc.ufscar.br, {gracan,taspardo}@icmc.usp.br

Summary. *This paper introduces CSTNews, a discourse-annotated corpus for fostering research on single and multi-document summarization. The corpus comprises 50 clusters of news texts in Brazilian Portuguese and some related material, which includes a set of single-document manual summaries and a set of multi-document manual and automatic summaries. The texts are annotated in different ways for discourse organization, following both the Rhetorical Structure Theory and Cross-document Structure Theory. The corpus is a result delivered within the context of the SUCINTO Project, which aims at investigating summarization strategies and developing tools and resources for that purpose. The design of the discourse annotation tasks and the decisions that have been taken during the annotation process are detailed in this paper.*

1. Introduction

Automatic text Summarization (AS) is the task of automatically creating a shorter version of one or more texts (Mani, 2001). It is a consensus that summaries are useful for several daily activities, such as selecting books and papers to read, getting updated with the latest episodes of some TV show, grasping the main ideas of some political discussion reported in newspapers, among others. The Web, in particular, has

Anais do III Workshop "A RST e os Estudos do Texto", páginas 88–105, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

contributed to increase interest in automatic summarization. Users are usually overloaded with news information and barely have enough time to digest them in their full form, e.g., the frequent tragedies due to climate changes or facts and rumors about some new pop idol. The types of information provided by Google Trends[1] or Google News[2], which demand specialized search engines, are incredibly useful. Besides, they present a challengeable environment for dealing with a huge amount of information, hence for summarizing tools.

Demands for AS have been massively signaled by the increasing modalities in which it happens. From the traditional single-document to the more recent multi-document summarization tasks, one may find the so-called update summarization, meeting summarization, cross-language/multilingual summarization, opinion summarization, e-mail and blog summarization, multimedia summarization, etc.

The ongoing SUCINTO project[3] tackles some of the modalities above. It aims at investigating and exploring generic and topic-focused multi-document summarization strategies for more feasible and intelligent access to on-line information provided by news agencies. This commitment brings together old and well-known scientific challenges from the first studies in summarization (back to the 50s) and several new and exciting challenges, e.g., to deal with redundant, complementary and contradictory information (which constitute the main multi-document phenomena), to normalize different writing styles and referring expression choices, to balance different perspectives of the same events and facts, to properly deal with evolving events and their narration in different occasions, and to arrange information pieces from different texts to produce coherent and cohesive summaries. An ultimate goal of this project is to pull the developed tools together as on-line applications in the Web for final users.

SUCINTO takes into consideration not only classical approaches to single and multi-document summarization, but also novel ones. Different paradigms for AS are explored, and knowledge of varied kinds is used, ranging from empirical and statistical ones to semantic and discourse models. Research interests include (i) the modeling of the summarization process (content selection, planning, aggregation, generalization, substitution, information ordering, etc.) by means of Cross-document Structure Theory (CST) (Radev, 2000), Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), ontologies, and language and summarization statistical models, (ii) the investigation of related tasks, namely, discourse parsing, topic detection, temporal annotation and resolution, coreference resolution, text-summary alignment, and multilingual processing, and (iii) the linguistic characterization of multi-document summaries and their manual production.

The project is mainly corpus-driven, i.e., tools and applications are drawn upon annotated corpus. This motivates our interest in having a rich, reference corpus, named the CSTNews corpus, first described by Aleixo and Pardo (2008a). It is composed of 50 clusters of news texts in Brazilian Portuguese (BP), with each cluster comprising 2 or 3 texts collected from on-line Brazilian news agencies as *Folha de São Paulo*, *Estadão*, *O Globo*, *Gazeta do Povo*, and *Jornal do Brasil*. Besides the original texts, each cluster conveys single-document manual summaries and multi-document manual and automatic summaries. The corpus is manually annotated in different ways for discourse

[1] http://www.google.com/trends

[2] http://www.google.com/news

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

organization, following both the Rhetorical Structure Theory and Cross-document Structure Theory.

RST and CST annotations are particularly helpful for content selection for the automatic production of summaries and have already been used by several authors in the area (see, e.g., Marcu, 2000a; Zhang et al., 2002; Seno and Rino, 2005; Carbonel et al., 2006; Afantenos et al., 2008; Uzêda et al., 2010; Jorge and Pardo, 2010; Jorge et al., 2011). While RST represents the relations among propositions inside a text and discriminates nuclear and satellite information, CST addresses the relationships among spans from several texts on the same topic. From the former structures, it is possible to have a relevance model to distinguish propositions that may be more relevant than others for summarization, taking into consideration the RST relations and their nuclearity. From the latter, it is possible to pinpoint redundant segments to recognize relevant information. Additionally, in doing so, it is feasible to keep only non-redundant information in a summary of a group of texts as well as to deal with the majority of the multi-document phenomena.

So far, RST and CST have not been used together for summarization. The main reason may be that there are not known resources that simultaneously have both reliable RST and CST-annotated texts. In literature we may find corpora that are annotated with only one of these models. There are some well-known corpora manually annotated with RST. For instance, there are the RST Discourse Treebank (Carlson et al., 2003), the Discourse Relations Reference Corpus (Taboada and Renkema, 2008), and the Penn Discourse Treebank (Prasad et al., 2008) for English; the Potsdam Commentary Corpus (Stede, 2004) for German; the CorpusTCC (Pardo and Nunes, 2004; Pardo, 2005), Rhetalho (Pardo and Seno, 2005), and Summ-it (Collovini et al., 2007) for Portuguese; the RST Spanish Treebank (da Cunha et al., 2011) for Spanish. For CST, there is only one known annotated corpus for English: the CSTBank (Zhang et al., 2002).

To the best of our knowledge, SUCINTO is the first project that attempts to build an annotated corpus based upon both theories, aiming at using such knowledge together for summarization. This paper focuses only in the first task, namely, how to proceed to a manual discourse annotation of the CSTNews corpus, in order to produce a reference corpus for AS. It first reports on the main annotation decisions, and then it describes the annotation task itself, followed by the achieved results. More details about the SUCINTO project are available in the project webpage.

Next section briefly introduces RST and CST, while the discourse annotation is reported in Section 3. Some final remarks are made in Section 4.

2. RST and CST

The Rhetorical Structure Theory (RST) was proposed by Mann and Thompson (1987) as a theory of text organization based upon its underlying propositions and their functions. More specifically, the theory prescribes a way to retrieve and generate the relationships among propositions under the assumption that the writer rhetorically organizes a text based upon his/her intentions towards the reader. Propositions express basic meaningful units, which are usually expressed by clauses or sentences at the surface of a text. The relationships are traditionally structured in a tree-like form (where larger units – composed of more than one proposition – are also related in the higher levels of the tree), although some recent works have argued that graphs are more

suitable than trees for text organization (see, e.g., Wolf and Gibson, 2006). Table 1 shows the original RST relations set.

Table 1 – Original RST relations defined by Mann and Thompson (1987)

Circumstance	Volitional Cause	Otherwise
Solutionhood	Non-Volitional Cause	Interpretation
Elaboration	Volitional Result	Evaluation
Background	Non-Volitional Result	Restatement
Enablement	Purpose	Summary
Motivation	Antithesis	Sequence
Evidence	Concession	Contrast
Justify	Condition	Joint

The Example (1) illustrates two clauses (numbered) whose corresponding propositions are in a Concession relation (Mann and Thompson, 1987, p. 13):

(1) [Although it is toxic to certain animals,]$_1$ [evidence is lacking that it has any serious long-term effect on human beings.]$_2$

New important relations were soon added to the original RST list, such as Means and List. Other relation sets have also been derived from the original one, usually based upon diversified purposes for text organization to address particular text genres and domains (see, e.g., Marcu, 1997). Following those, Pardo (2005) also modified that relation set by adding some of the Marcu's relations, resulting in the DiZer relation set shown in Table 2.

Table 2 – Relation set defined by Pardo (2005)

Circumstance	Volitional Cause	Otherwise	Means
Solutionhood	Non-Volitional Cause	Interpretation	List
Elaboration	Volitional Result	Evaluation	Explanation
Background	Non-Volitional Result	Restatement	Comparison
Enablement	Purpose	Summary	Conclusion
Motivation	Antithesis	Sequence	Attribution
Evidence	Concession	Contrast	Parenthetical
Justify	Condition	Joint	Same-Unit

It is noticeable that, differently from the original RST relations, some relations defined by Marcu are purely structural, in that they do not address discourse relationships themselves, but signal how constituent text spans are connected at the surface instead. Parenthetical and Same-Unit relations are examples of those relations. The former signals additional details that are usually introduced in a bracketed way in the text; the latter pinpoints non-adjacent text segments that only together convey a full single proposition.

RST also defines what is called nuclearity for each relation. The propositions in a relation are classified as nuclei (i.e., more important propositions) or satellites (i.e., complementary information), and this classification reflects the author's intention. Relations with one nucleus and one satellite are said to be mononuclear relations. Relations that only have nuclei (where all the propositions are equally important) are said to be multinuclear relations. Sequence, Contrast, List, Joint and Same-Unit are multinuclear relations; the others are mononuclear relations.

Figure 1 shows an example of a complete RST structure that embeds the text exemplified in (1). It has been produced by using the RSTTool (O'Donnell, 2000)[4]. In mononuclear relations, the direction of an arrow is from a satellite towards its corresponding nucleus (which is also signaled by a vertical line). So, the nucleus of the Concession relation below is segment 3, whilst its satellite is segment 2. The horizontal line above such relation indicates that there is an RST subtree comprising both segments 2-3. In addition, this subtree is the entire satellite of the Elaboration relation, whose nucleus is segment 1. The full tree for such a text is indicated again by a horizontal line signaled by 1-3 above it. One may notice that the result of organizing a text based upon RST is a hierarchical structure, and leaves are text spans supposed to correspond to the propositions, or, as named by Marcu (1997), the underlying elementary discourse units (EDUs).

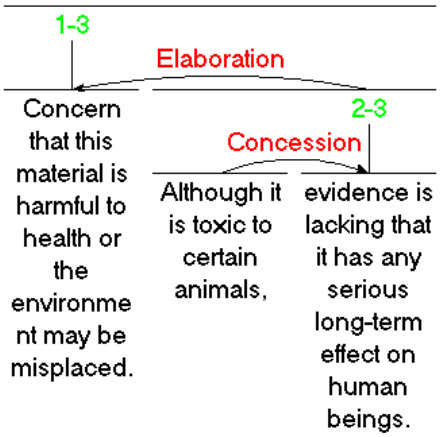

Figure 1 – Example of an RST structure

One of the main difficulties in analyzing a text under RST assumptions is that such a task is highly dependent on the understanding of the text by the RST expert. Often humans analyzing the same text may disagree in several aspects, ranging from the identification of the EDUs to the definition of the relations that may hold amongst them and their nuclearity. The first problem refers to text segmentation; the second to grasping adequate discourse relations, or, more importantly, to grasping the writer's intentions that are subjacent to surface choices. In fact, it is pretty acceptable that more than one RST structure may hold for the same text.

Inspired by RST and related work (Trigg, 1983; Trigg and Weiser, 1986; Radev and McKeown, 1998), the Cross-document Structure Theory (CST) was proposed by Radev (2000) as a way of relating text passages from different texts on the same topic. Diverse news about the same event, e.g., the earthquakes in Japan, published on-line, are examples of that. Therefore, differently from RST, CST was devised mainly for dealing with multi-document organization, and may be used to solve several problems, such as summarization and question-answering ones. It may provide the means for a more intelligent information processing, particularly if we consider that it allows for dealing with redundancy and other different multi-document phenomena conveyed by a group of texts with its set of 24 original relations (Table 3).

[4] http://www.wagsoft.com/RSTTool/

Table 3 – Original CST relations

Identity	Modality	Judgment
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Contrast
Historical background	Indirect speech	Parallel
Cross-reference	Refinement	Generalization
Citation	Agreement	Change of perspective

Such relations do not connect every text passages, but only those that are more closely related. In other words, the relations are commonly identified among pairs of sentences, coming from different sources, which are related by a lexical similarity significantly higher than random. According to Radev, CST relations may link together segments of any degree of granularity, ranging from words and clauses to sentences, paragraphs, or even entire texts. The result of annotating a group of texts is a graph (instead of a tree), which may be probably disconnected, since not all segments present relations with other segments. Considering sentences as the text segments in the discourse analysis, Example (2) shows the Equivalence relation between two sentences from different texts (Radev, 2000, p. 79). This relation states that the two text sentences, (a) and (b), from different sources, have similar content.

(2) a. Ford's program will be launched in the United States in April and globally within 12 months.
b. Ford plans to introduce the program first for its employees in the United States, then expand it for workers abroad.

Instead of pinpointing nuclearity between text spans, CST differentiates symmetric and non-symmetric relations. Symmetric ones are non-directed, like the Equivalence relation that holds between the sentences (a) and (b) in Example (2). Being unordered, it is possible to put the sentences together, or read them, in any order. Subsumption, on the other hand, is asymmetric, thus it is signaled by a directed arrow. In this case, one sentence contains more information than another. The opposite is not true. In the following Example (3), extracted from Radev (2000, p. 80), sentence (b) subsumes (a) because the crucial information in (a) is also included in (b), which presents additional content: “the court”, “last August”, and “sentenced him to life”.

(3) a. John Doe was found guilty of the murder.
b. The court found John Doe guilty of the murder of Jane Doe last August and sentenced him to life.

Some researchers have also changed the original relation set. Zhang et al. (2003), for example, refine the relation set and end up with 18 relations. Based on a corpus of news texts, Maziero et al. (2010) also shortened the original set, ending up with 14 relations (shown in Figure 2). The authors not only end up with 14 relations, but redefined all of them and proposed a typology of relations. According to the typology (Figure 2) (Maziero et al., 2010, p. 6), “content” relations address mainly the meaning of the corresponding passages, whilst “presentation/form” relations address mainly writing styles and authorship information. The authors claim that it is not possible to have more

than one content relation between the same information in the text segments. They also notice that presentation/form relations usually come along with content relations.

Zhang et al. claimed (and so did Maziero et al.) that CST relations are unlikely to hold between segments that are lexically very dissimilar to each other. Although this is not always true, it allows for constraining the annotation process to a reasonable amount of relations that hold for groups of texts, since empirical evidence has shown that humans are capable of making sense of very dissimilar and different sentences, resulting in a great number of CST relations. Even for short texts, e.g., two texts of 10 sentences each, humans would be able to relate every possible sentence pair, which would easily reach 100 or more relations. Considering that a sentence pair may have more than one CST relation, this scenario gets even worse and unmanageable.

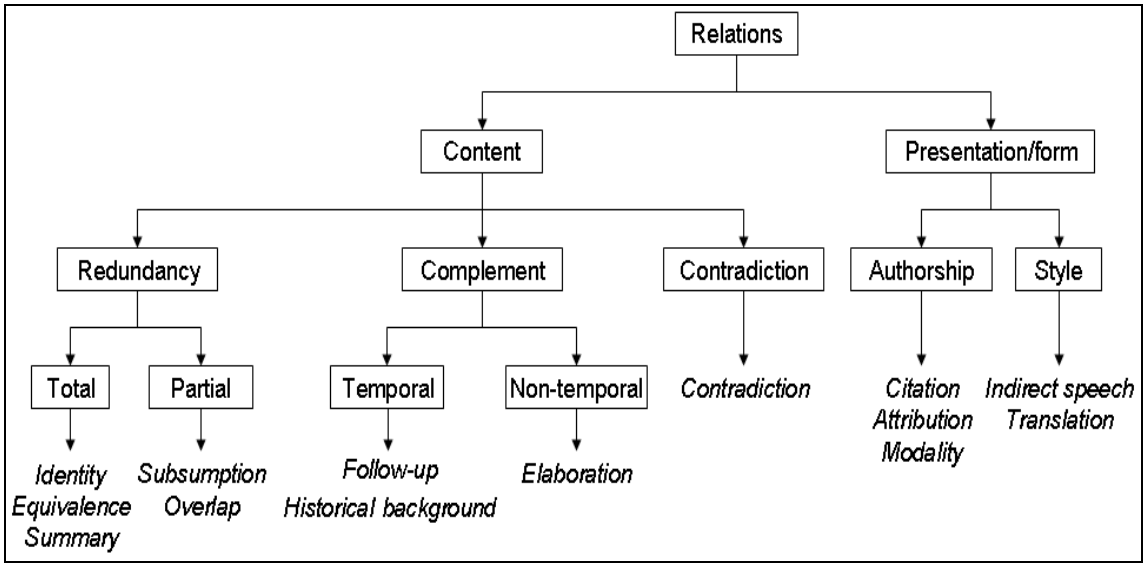

Figure 2 – Typology of CST relations

Although CST seems simpler than RST, it involves very difficult issues concerning its set of relations. Besides this possibility of relating every segment pair, ambiguity often takes place, which may be due to different text interpretations or to subspecified information in the texts (for example, when the publication dates of news are not specified in several newspapers, it is difficult to determine the appropriate order to reproduce some events).

3. Discourse annotation of the CSTNews Corpus

The CSTNews corpus is composed of 50 clusters of news texts collected in 2007. They address several topics from popular on-line news agencies in Brazil, namely, *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*. Each cluster conveys 2 to 3 texts written in BP, collected according to their repercussion by the time they were published.

Figure 3 shows the distribution of clusters by categories. The corpus sums up 140 texts altogether, amounting to 2,088 sentences and 47,240 words. In average, the corpus conveys 2.8 texts, 41.76 sentences and 944.8 words per cluster.

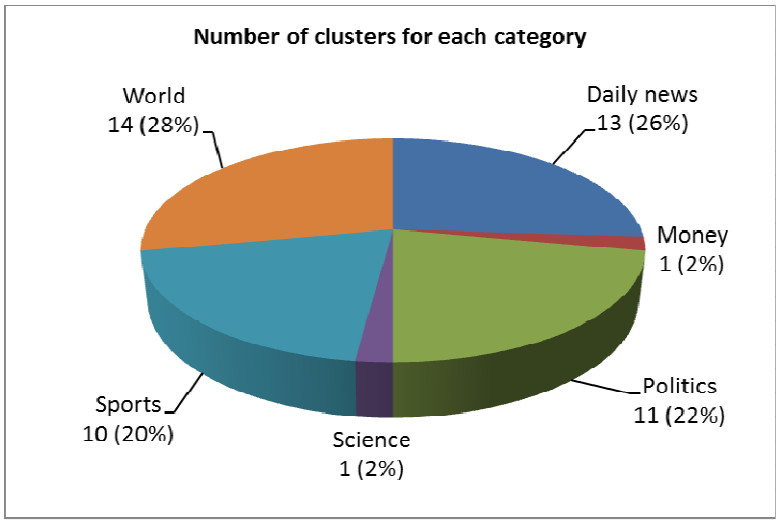

Figure 3 – Distribution of cluster by categories

The next subsections introduce the RST and the CST annotation of this corpus, respectively.

3.1. RST annotation

RST annotation has been performed by 8 annotators, 4 of them with deep knowledge on RST and some experience on annotation. They all went through a training phase first, in which a study of the theory was carried out based mainly on the technical report by Mann and Thompson (1987) and the reference manual of Carlson and Marcu (2001). This material has been proved applicable to texts in Portuguese (Seno and Rino, 2005).

Segmentation and annotation training took approximately 2 weeks each. During this process, some segmentation rules were adapted to deal with the BP language. For instance, all relative clauses should be segmented, instead of segmenting only non-restrictive ones. The reason for this adaptation relies in the difficulty of distinguishing restrictive and non-restrictive clauses (also known as defining and non-defining clauses) for some cases in BP. Also, clausal complements of attribution verbs should only be segmented if the corresponding subject was mentioned and animated (e.g., a person, a group of people, organizations or institutions, when those referred to groups of people). Inner speech clauses and authored texts should also be segmented: even literal transcriptions are segmented when they involve more than one clause. Very often texts in Portuguese convey more complex sentences than texts in English. As a result, syntactical realizations that involve, e.g., relative or elliptical constructions are quite common. Specific segmentation rules have been defined for those. 'Segment conjoined phrases even if the subject is implicit or the verb is elliptical.' is an example of segmentation rule. The occurrence of strong discourse markers may also indicate the segmentation sometimes.

All the segmentation rules are shown in Table 4. Examples come along their English translations. Segment boundaries are indicated by the symbol 'l'.

Table 4 – Segmentation Rules for Brazilian Portuguese

Rule #	Rule description	Examples
1	Segment sentences that end with ´.´, ´!´, or ´?´	É uma final. | *This is a final game.* |
		Qual é o programa mais importante de seu computador? | *What is the most important program of your computer?* |
2	Segment full phrases (verbs explicit) or text segments with no verb, but signaled by strong discourse markers.	Eu não sabia das estatísticas, | mas acredito nas pessoas da Uefa. | *I didn´t know of such statistics,* | *but I trust people from Uefa.* |
		A partir da temporada 2012/2013, | a Alemanha terá quatro vagas na Copa dos Campeões. | *From the 2012/2013 games onwards,* | *Germany will have four places in the Champions Cup.* |
3	Segment phrases inside authored texts.	"Torço pelo Rodrigão, | que é meu amigo, | mas todos são merecedores." | *"I go for Rodrigão,* | *who is my friend,* | *but all of them deserve it."* |
4	Do not segment subject or object phrasal complements.	É muito raro alguém perder o jogo e arrancar a classificação no campo do adversário. | *It is quite uncommon one loosing the game and getting classified in the competitors' field.* |
5	Segment conjoined phrases even if the subject is implicit or the verb is elliptical.	Um dos macacos agarrou um livro de fotos | e começou a olhar as imagens. | *One monkey grasped the photo album* | *and began looking at the images.* |
		70% dos recursos são pagos pelo Estado | e os 30% restantes, pelas prefeituras regionais. | *70% of the incomes are paid by the state* | *and the remaining 30%, by the town halls.* |
6	Segment participial phrases only if they are explicitly marked (e.g., by commas).	Eliminado do "BBB 11" no domingo passado, | Maurício se manteve fiel ao seu grupo de amigos. | *Excluded from "BBB 11" last Sunday,* | *Maurício kept close to his group of friends.* |
		O navegador da MS chegou a 96% do mercado, | esmagando o finado Netscape Navigator. | *The MS navigator reached 96% of the market,* | *crushing the already-dead Netscape Navigator.* |
7	Segment every relative clause, either restrictive, or explicative.	Ela é uma menina | que sonhou. | *She is a girl* | *that has dreamed.* |
		Graham Waspe, | que perdeu a visão de um olho, | disse que | não queria desistir... | Graham Waspe, | who lost an eyesight, | said that | he

		would not given up. |
8	Segment when there are attributive or "public" verbs, since mentioned and animate subjects (either a person or a group of people, or an institution taking the place of a group of people), or when there are strong markers of Attribution. In any case there must be a verb in the main clause.	Maurício disse à Folha que | sua torcida é para Rodrigão. | *Maurício said to Folha that | he would go for Rodrigão. |*
		De acordo com o governo, | as santas casas respondem por 55% das internações. | *According to the state, | the public hospitals answer for 55% of the hospitalizations. |*
9	Segment every text span that refers to parenthetical information, usually marked with graphic signs such as parenthesis, hyphens, colon, etc.	A resposta mais provável: | o navegador da internet. | *The answer more feasible: | the Internet navigator. |*

In addition to adapting segmentation rules, discourse markers in Portuguese were also classified in *strong* or *weak*, according to how clearly they indicate the discourse structure and, therefore, some discourse relations. Strong markers include, for example, "porque" (because), "por meio de" (through, by means of), "além de" (also, besides), "quando" (when), "se" (if), "durante" (during), "após" (then, after), "mas" (but, besides), and "como" (for example, like). Weak markers are, for example, "e" (and), "com" (with), and "em" (in). Those discourse markers sets are not exhaustive because there are many cases that are made clear only in a proper context.

It is important to say that all the problems that where point out in this work are not particular from Portuguese language. Any intended RST analysis for other languages may require some adaptations according to its own characteristics, in order to do a proper annotation.

After training, the annotation started. Two months were necessary for the conclusion of the task. For both, segmentation and annotation, RSTTool (O'Donnell, 2000) has been used, which allows for a semi-automatic RST annotation. The relation set shown in Table 2 was adopted.

The annotation was performed 5 days a week in one-hour meetings, following five steps as described next:

1. Ideally, one cluster is annotated per meeting.
2. Groups of two or three analysts annotate each text of the cluster. Ideally, up to 3 texts in a cluster might be considered a day, since there are 8 RST analysts performing the task.
3. Those groups must be reorganized each meeting, in order to avoid bias.
4. When annotators of a group do not arrive at a consensus, they present the problem to the other analysts in order to disentangle it. If there is still no solution, a generic strategy is chosen. For example, if an RST relation choice is unclear, the most generic one is preferable.
5. Every 10 clusters annotated, one meeting is driven towards annotation agreement. In this case, all the groups segment and annotate the same text, in a two-phase basis. First, each group issues its segmentation data to the annotation session coordinator, in order to save data produced with no intervening. Then, all the experts discuss, and agree upon the segmentation results. Modifications are allowed for agreement, in order to proceed to the RST analysis. Differently from

segmentation, RST trees are not verified for agreement. Instead, they are automatically computed for agreement statistics.

In general, the annotators tried to perform the RST analysis in an incremental way, in order to take advantage of the text organization produced by its writer. This analysis assumes that adjacent clauses inside sentences must be related first. Then, adjacent sentences inside paragraphs must be related. Finally, adjacent paragraphs are related. Since language use is almost unrestricted and several writing styles exist, such incremental analysis may not always apply, but it is undoubtedly an interesting analysis heuristic and is also useful for disambiguating possible analyses.

It was also very interesting to observe the different analysis styles used by the annotators. While some annotators proceeded to complete text segmentation before choosing the relations, others performed intuitive topic segmentation, segmenting and relating the segments of each topic before joining these topic blocks in the highest levels of the RST tree.

Figure 4 shows the number of occurrences of each relation in the corpus. Some relations were very frequent (the more generic ones, as Elaboration and List), while others were rare or never happened (as Summary and Otherwise). It is also interesting to notice the large numbers of structural relations, such as Parenthetical and Same-unit.

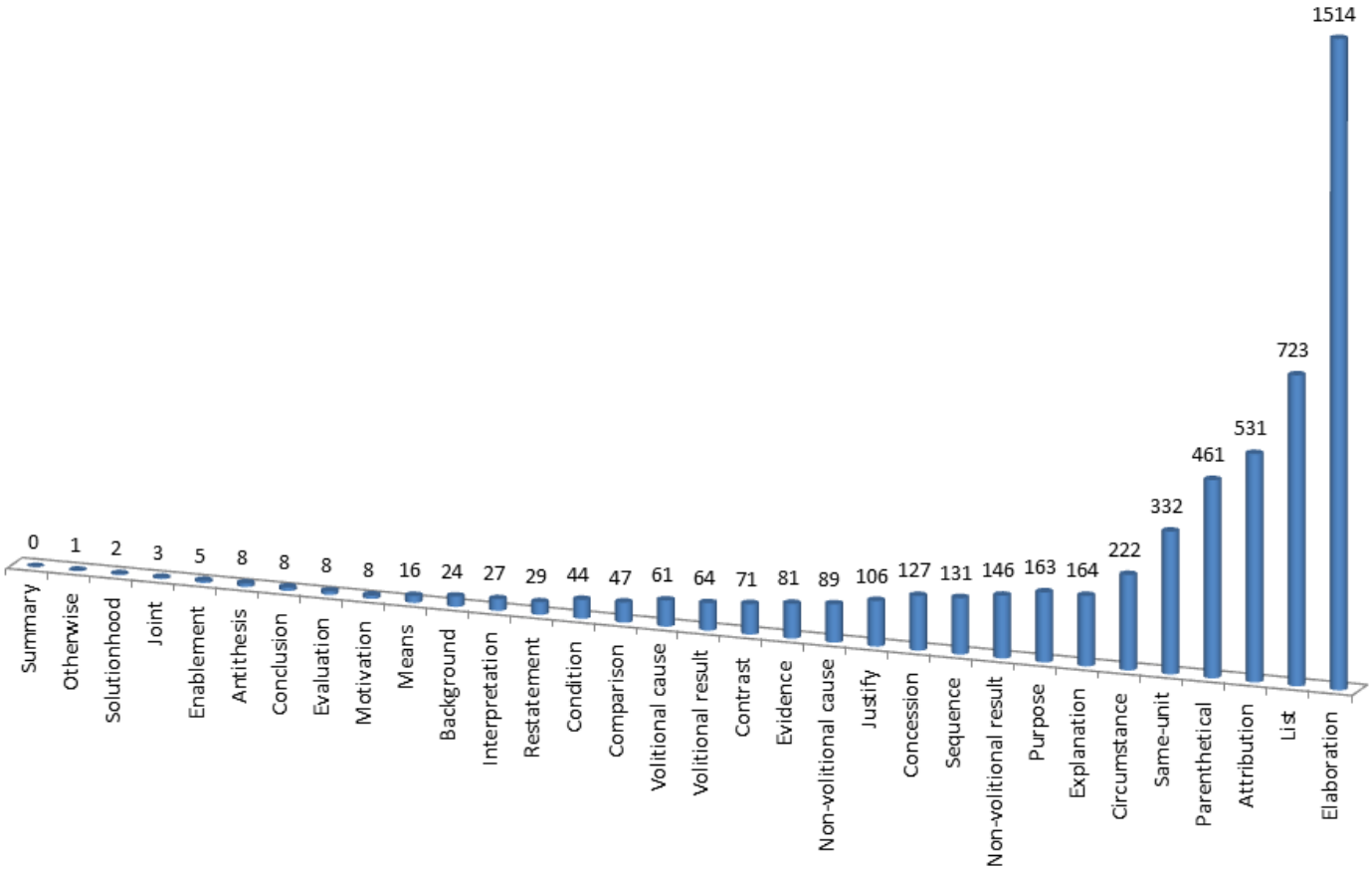

Figure 4 – RST relations in the corpus

Agreement between annotators has been automatically computed by using the tool RSTeval[5] (Maziero and Pardo, 2009), which had its basis in the work of Marcu (2000b). The methodology under RSTeval has tried to avoid the subjectivity of the RST analysis by using a systematic and deterministic procedure to compare rhetorical trees. Accordingly, in this tool, given a set of RST trees, a tree must be elected as the ideal one and the others are compared to it based on 4 criteria:

1. Simple textual segments;
2. Complex textual segments (i.e., two or more segments related by some RST relation(s));
3. Nuclearity of every text segment;
4. RST relation that holds between the segments.

The well-known Precision (P), Recall (R) and F-measure (F) metrics are computed by RSTeval for each RST tree and express the degree of similarity among trees of a same text. Precision represents the number of correct elements (*C*) (e.g., simple or complex textual segments, nuclearity or relations) of any RST tree *T,* divided by the total number of elements of *T* (Formula 1). Recall is the proportion of correct elements of *T,* regarding the total number of elements of an ideal tree *IT* (Formula 2). F-measure represents the harmonic mean between Precision and Recall (Formula 3).

$$P = C / |T| \quad (1)$$

$$R = C / |IT| \quad (2)$$

$$F = (2*R*P) / (R+P) \quad (3)$$

In order to illustrate the annotation agreement process, let us consider a cluster composed by four RST trees. Firstly, a tree is selected as the ideal one and the others 3 are compared to it, considering the 4 criteria mentioned before. This process is repeated four times, so that each time a different tree is selected as optimal. Then, the average agreement values are calculated for each criterion. Table 5 shows these values for the corpus. Precision and recall are the same because all the annotations are compared to one another.

Table 5 – Precision, Recall and F-measure average values

Evaluated Criteria	Precision	Recall	F-Measure
Simple Textual Segments	0.91	0.91	0.91
Complex Textual Segments	0.78	0.78	0.78
Nuclearity	0.78	0.78	0.78
Relations	0.66	0.66	0.66

According to the results, the best values of agreement were achieved in the segmentation process (simple textual segments), computed before the annotators discuss about it. This is mainly due to the segmentation rules that make this task less subjective than the others. As expected, the worst agreement values were obtained for the relations the annotators indicated.

[5] http://www.nilc.icmc.usp.br/rsteval/

For comparison purposes, using his original (similar) evaluation strategy, Marcu (2000b) reports numbers for a group of 5 texts annotated by 2 humans. He got the following results: 0.88 precision and recall for simple textual segments; 0.90 precision and recall for complex textual units; 0.79 precision and 0.88 recall for nuclearity; and 0.83 precision and recall for relations. da Cunha et al. (2011) also use the same method for evaluating the agreement in the annotation of the RST Spanish Treebank. Applied to 84 texts annotated by 10 humans, the results were: 0.87 precision and 0.91 recall for simple textual segments; 0.86 and 0.87 for complex textual segments; 0.82 precision and 0.85 recall for nuclearity; 0.77 precision and 0.78 recall for relations.

Although the texts, languages and the amount of data used by these other authors and in this work are very different (therefore, not allowing a fair comparison between the works), such comparison gives an idea of the human ability to agree on the RST annotation process.

In general, we consider that the agreement results in this work were quite satisfactory, given the subjectivity of the task.

3.2. CST annotation

Four computational linguists were in charge of annotating the corpus according to CST. In fact, the work reported by Maziero et al. (2010) and the CST typology shown in Figure 2 are some of the results of this annotation.

Similarly to the RST annotation, the CST annotation was also preceded by training. For 3 months a group of 4 annotators have explored the theory and some news texts have been annotated. A discussion on the analyses was thus carried out and, as result of this training process, it was possible to refine the CST relation set. The refinement was carried out by (i) removing a few relations that were not conveyed in the training texts and were not expected to happen for the CSTNews texts we were working on, and by (ii) merging some relations. The relations were merged when they could not be differentiated by the annotators. For example, Refinement, Description and Elaboration relations were merged into Elaboration. Our final relation set is shown in Figure 2, which contains 14 relations.

Besides the refinement of the relation set, more formal definitions were also established for each relation. The definition for the Subsumption relation is illustrated in Figure 5. One may see that a relation definition is composed of 5 fields: the relation name, its type (according to the path in the typology that takes to the relation), its directionality (when it is the case), constraints for the relation to apply, and any (optional) additional comments that might help to understand and to use the relation.

Relation name: Subsumption
Type: Content → Redundancy → Partial
Directionality: S1 → S2
Restrictions: S1 presents the information in S2 and some additional information
Comments: S1 presents some content X and Y, S2 presents only X

Figure 5 – Definition for Subsumption relation

The annotation itself has taken about 2 months in a daily one-hour meeting basis. Each day was enough for annotating 1 cluster, since each possible text pair inside a cluster was annotated by a different annotator. In special annotation sessions, usually once a

week, all the annotators annotated the same texts in order to compute agreement and to measure how well understood the CST annotation was.

For the annotation, it was used the CSTTool (Aleixo and Pardo, 2008b), a tool for CST semi-automatic annotation. CSTTool performs 2 tasks. At first, the tool automatically segments the input texts into sentences, which are the basic segments used in the corpus annotation under consideration. Then, using the traditional word overlap measure, it computes the lexical similarity for all sentence pairs (considering that the sentences come from different texts) and indicates the better sentence pairs to the annotator, who, in turn, manually selects the most appropriate relation(s) to hold between the sentences.

The word overlap measure between 2 sentences S1 and S2 is computed by using Formula 4.

$$\text{word overlap} = \frac{\text{number of common words in S1 and S2}}{\text{number of words in S1} + \text{number of words in S2}} \quad (4)$$

This formula produces a number between 0 and 1, with 0 indicating that the sentences do not have any word in common. Following other works in the area (Zhang and Radev, 2004 – for English; Aleixo and Pardo, 2008c – for Portuguese), we adopted 0.12 as the minimum value to consider that a sentence pair might have CST relations. It is important to notice that the CSTTool uses such value only to indicate candidate sentence pairs to the annotator, not to avoid the annotator to annotate other sentence pairs s/he might find important. The annotator has access to every possible sentence pairs from the texts.

Figure 6, extracted from Maziero et al. (2010, p. 7), shows the number of each relation in the corpus. It is interesting to notice that some relations were very frequent (e.g., Elaboration and Overlap relations), while others were not (e.g., Citation, Modality and Translation relations – in fact, the Citation relation never occurred).

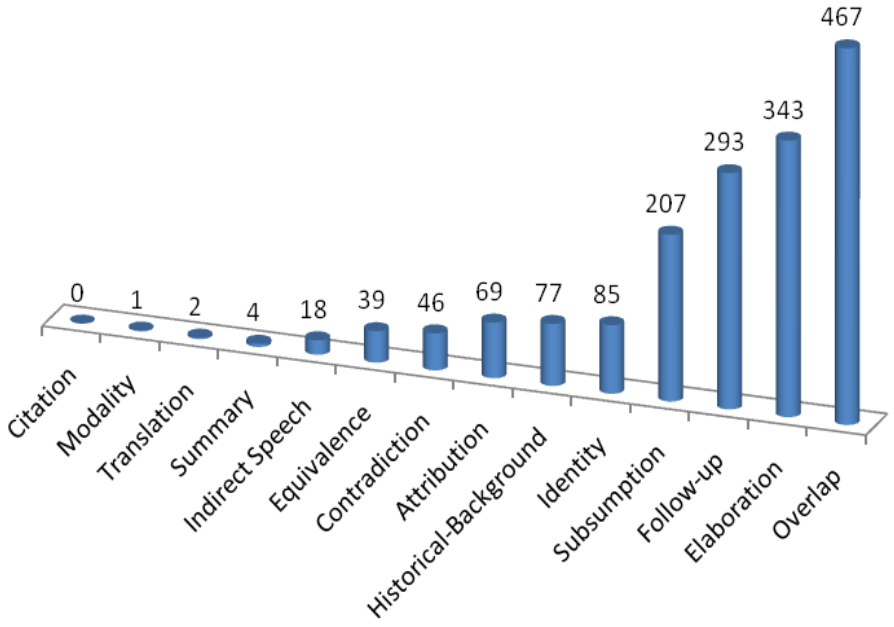

Figure 6 – CST relations in the corpus

For computing the annotation agreement, we used two measures. Firstly, the already traditional kappa measure (Carletta, 1996) was used to compute agreement for the relations the annotators indicated, for the directionality of these relations, and, finally, to the type of relations that were used (instead of the relations themselves – this might indicate that, although some relation might be difficult to identify, the annotators could still have some idea of what type of relation might hold, according to the typology in Figure 2). Although it highly depends on the task that is under evaluation, a kappa value of 0.6 is usually accepted as the minimum value for which the annotation may be considered reliable. For discourse annotation tasks like the one that was performed here, it is natural to expect a lower minimum value.

Table 7 shows the kappa agreement numbers for the CST annotation. One may see that the results were quite good considering how difficult the task is. As expected, when the relation types are used, results are better.

Table 7 – Kappa agreement for the CST annotation

Agreement parameters	Agreement value
Relations	0.50
Directionality	0.44
Relation types	0.61

The other agreement measure that was used was the percent agreement, which is based on the number of times that all annotators indicated the same (in this case, computing the full agreement), the majority of the annotators indicated the same (partial agreement), or none of the previous (null agreement). Such measure was used in order to compare our results to the only other result in CST annotation that is found in the literature (Zhang et al., 2002 – for English). Although the works use different corpora for different language, such comparison may give an idea of the state of the art in the area.

Table 8 shows the percent agreement results for the CST annotation. Zhang et al. point out 58% of full and partial agreement (computed together) for the relations, while here it was obtained more than 80% of agreement. Zhang et al. did not use the kappa measure.

Table 8 – Percent agreement (in %) for the CST annotation

Agreement parameters	Full agreement	Partial agreement	Null agreement
Relations	54	27	18
Directionality	58	27	14
Relation types	70	21	9

For the relation types, the very good 91% agreement was obtained.

Such agreement numbers not only shows that the annotation process was well conducted, but that the results are reliable enough to be used in future researches.

4. Final remarks

This paper described in detail the discourse annotation of the CSTNews corpus, which aims at supporting the investigation of deep strategies on single and multi-document summarization for Brazilian Portuguese texts. Besides the subjectivity of RST and CST,

the annotation experience showed that it is possible to obtain some level of systematization of the task, which allows reaching acceptable levels of agreement.

The corpus, tools and resources that were developed in this work are all available on-line for use by the research community. We hope the CSTNews corpus may foster research not only on summarization and discourse analysis, but also in other Natural Language Processing areas.

Future work includes extending the corpus with other levels of annotation, as identification and normalization of temporal expressions and resolution of co-references.

Acknowledgments

The authors are grateful to FAPESP, CAPES and CNPq for supporting this work.

References

Afantenos, S.D.; Karkaletsis, V.; Stamatopoulos, P.; Halatsis, C. (2008). Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. *Journal of Intelligent Information Systems*, Vol. 30, N. 3, pp. 183-226

Aleixo, P. and Pardo, T.A.S. (2008a). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva CST (Cross-Document Structure Theory).* Technical Report, N. 326. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP.

Aleixo, P. and Pardo, T.A.S. (2008b). *CSTTool:* um parser multidocumento automático para o Português do Brasil. In the P*roceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence*. Salvador-BA.

Aleixo, P. and Pardo, T.A.S. (2008c). Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana*, pp. 298-303. Vila Velha-ES.

Carbonel, T.I.; Seno, E.R.M.; Pardo, T.A.S.; Coelho, J.C.; Collovini, S.; Rino, L.H.M.; Vieira, R. (2006). A Two-Step Summarizer of Brazilian Portuguese Texts. In the *Proceedings of the 4th Workshop on Information and Human Language Technology*. Ribeirão Preto-SP.

Carletta, J. (1996). *Assessing Agreement on Classification Tasks: the Kappa Statistic*. *Computational Linguistics*, Vol. 22, N. 2, pp. 249-254.

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

Carlson, L; Marcu, D.; Okurowski, M.E. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J.V. Kuppevelt and R. Smith (eds.), *Current Directions in Discourse and Dialogue*, pp. 85-112. Kluwer Academic Publishers.

Collovini, S.; Carbonel, T.I.; Fuchs, J.T.; Coelho, J.C.B.; Rino, L.H.M.; Vieira, R. (2007). Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In the *Proceedings of the 5th Workshop in Information and Human Language Technology*. Rio de Janeiro-RJ.

da Cunha, I.; Torres-Moreno, J-M.; Sierra, G. (2011). On the Development of the RST Spanish Treebank. In the *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 1-10. Portland-Oregon.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82. Uppsala/Sweden.

Jorge, M.L.R.C.; Agostini, V.; Pardo, T.A.S. (2011). Multi-document Summarization Using Complex and Rich Features. In *Anais do VIII Encontro Nacional de Inteligência Artificial*. Natal-RN.

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. (2000a). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

Marcu, D. (2000b). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, Vol. 26, pp. 396-448.

Maziero, E.G. and Pardo, T.A.S. (2009). Automatização de um Método de Avaliação de Estruturas Retóricas. In the *Proceedings of the RST Brazilian Meeting*, pp. 1-9. São Carlos-SP.

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science*, pp. 60-69. Funchal/Madeira.

O'Donnell, M. (2000). RSTTool 2.4 -- A Markup Tool for Rhetorical Structure Theory. In the *Proceedings of the International Natural Language Generation Conference*, pp. 253-256. Mitzpe Ramon, Israel.

Pardo, T.A.S. and Nunes, M.G.V. (2004). *Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil*. Technical Report, N. 231. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 73p.

Pardo, T.A.S. and Seno, E.R.M. (2005). Rhetalho: um corpus de referência anotado retoricamente. In *Anais do V Encontro de Corpora*. São Carlos-SP.

Pardo, T.A.S. (2005). *Métodos para Análise Discursiva Automática*. PhD Thesis. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 211p.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.; Webber, B. (2008). The Penn Discourse Treebank 2.0. In the *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech/Morocco.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong-China.

Radev, D.R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, N. 3, pp. 469-500.

Seno, E.R.M. and Rino, L.H.M. (2005). Co-referential chaining for coherent summaries through retorical and linguistic modeling. In the *Proceedings of the Workshop on Crossing Barriers in Text Summarization Research/RANLP*. Borovets-Bulgaria.

Stede, M. (2004). The Potsdam Commentary Corpus. In the *Proceedings of the ACL Workshop on Discourse Annotation*.

Taboada, M. and Renkema, J. (2008). *Discourse Relations Reference Corpus*. Simon Fraser University and Tilburg University. Available at http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. PhD Thesis. University of Maryland, College Park MD.

Trigg, R., and Weiser, M. (1986). TEXTNET: A Network-Based Approach to Text Handling. *ACM Transactions on Office Information Systems*, Vol. 4, N. 1, pp. 1-23.

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2010). A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, Vol. 6, N. 4, pp. 1-20.

Wolf, F. and Gibson, E. (2006). *Coherence in Natural Language*. MIT Press.

Zhang, Z.; Blair-Goldensohn, S.; Radev, D.R. (2002). Towards CST-Enhanced Summarization. In the *Proceedings of AAAI Conference*. Edmonton-Alberta.

Zhang, Z.; Otterbacher, J.; Radev, D.R. (2003) Learning cross-document structural relationships using boosting. In the *Proccedings of the Twelfth International Conference on Information and Knowledge Management CIKM 2003*, pp. 124-130, New Orleans-Louisiana.

Zhang, Z. and Radev, D.R. (2004). Learning cross-document structural relationships using both labeled and unlabeled data. In the *Proceedings of the International Joint Conference on Natural Language Processing*. Hainan Island-China.