On the Development and Evaluation of a Brazilian Portuguese Discourse Parser

Thiago Alexandre Salgueiro Pardo[1]
Maria das Graças Volpe Nunes[1]

Resumo: Apresentamos neste artigo o processo de desenvolvimento e avaliação de um analisador discursivo automático para o português brasileiro. Seguindo a Teoria de Estruturação Retórica, o DiZer é um sistema simbólico baseado na ocorrência de marcadores textuais, fazendo uso de templates discursivos extraídos de um corpus de textos científicos para identificar a construir a estrutura discursiva de textos. A avaliação do DiZer mostra resultados satisfatórios para textos científicos e jornalísticos, apesar do sistema não ter sido delineado para o gênero jornalístico, o que demonstra a portabilidade do sistema.

Palavras-chave: análise discursiva automática, RST

Abstract: *We present in this paper the development process and the evaluation procedure of a Brazilian Portuguese discourse parser called DiZer. Based on Rhetorical Structure Theory, DiZer is a symbolic cue phrase-based analyzer that makes use of discourse templates learned from a corpus of scientific texts to identify and build the discourse structure of texts. DiZer evaluation shows satisfactory results for scientific and news texts, even tough it was not designed for the latter, which demonstrates DiZer portability.*

Keywords: *discourse parsing, RST*

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC). Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo. CP 668, 13.560-970, São Carlos-SP.
{taspardo,gracan@icmc.usp.br}

1 Introduction

It is well known that a text is more than just a simple sequence of juxtaposed sentences. It has a highly elaborated underlying discourse structure. In general, this structure represents how the pieces of information conveyed by text segments correlate and make sense together.

The ability to automatically derive discourse structures of texts is of great importance to many applications in Computational Linguistics. For instance, it may be very useful for identifying relevant information of a text to produce its summary (see, for instance, [23][32][38]), determining the possible antecedent text spans for a referred term in co-reference resolution [8][44], producing coherent texts according to communicative goals in text generation [28][29][42], adequately rearranging spans of texts being translated in a machine translation task [25], identifying missing important components in a essay scoring procedure [4], elaborating an answer to better satisfy an specific user in a question answering system [3], among other natural language applications.

There are several discourse theories that try to represent different aspects of discourse. Grosz and Sidner [13] propose how to model the intentions in a text and their relationships; Jordan [16] and Kehler [17] present semantic relations for structuring text content; Mann and Thompson [20] introduce the most prominent discourse theory in Computational Linguistics, the RST, for representing the rhetorical organization of texts, which is focused in this paper.

Some discourse parsers based on RST are available for both English and Japanese languages. Some of these will be reviewed in the next section. In this paper we present a discourse parser for Brazilian Portuguese, DiZer (DIscourse analyZER), that, to our knowledge, is the first one available for this language. The development of resources and tools for the computational processing of Brazilian Portuguese is quite recent if compared to English. Nonetheless it has already produced good part-of-speech taggers (e.g., [1]), syntactic parsers [2][49] and applications such as text summarizers [43] and writing tools [11]. A discourse parser represents then an important step towards more interesting and sophisticated language processing tools.

DiZer is a RST discourse parser that presents unique characteristics when compared to other parsing approaches. It is a symbolic system that makes use of discourse templates that codify the correspondence between discourse structuring and cue phrases that texts present, for example, the discourse makers. The templates amount to about 750 and were manually encoded from a comprehensive corpus analysis of Computer Science scientific texts. For being customized for this kind of text, DiZer also exploits genre dependent information, more specifically, indicative phrases and words, word classes, and heuristics. DiZer analysis strategy is said to be incremental, in the sense that it takes advantage of the fact that the text writer clusters related information and hierarchically organizes such information in clauses, sentences and paragraphs according to their relationship and importance in the text.

DiZer was evaluated for scientific and news texts, even though it was not designed for this last text genre. This evaluation with news texts was carried out to test the system

portability to other text genres, as well as the templates genre dependence. DiZer performed well for both genres.

Initially, in Section 2, we introduce RST and relevant related work on discourse parsing. In Section 3, we describe DiZer, its main modules and information repositories, as well as the process of corpus annotation and knowledge extraction to produce DiZer knowledge sources. Section 4 reports the efforts on building a rhetorically annotated reference corpus for DiZer evaluation and the system performance for scientific and news texts. Some conclusions and final remarks are presented in Section 5.

2 RST and discourse parsing

In this section, we first introduce the main concepts of RST [20] and, then, describe important aspects of related work on discourse parsing.

2.1 RST and the discourse levels

According to RST authors, rhetoric represents the text functional organization, i.e., the function of the text parts and how they are organized in order to achieve the text communicative goal, i.e., the intention the writer had in mind when prepared the text. Hovy [15] defines rhetoric as the "touchable" part of pragmatics.

RST states that all propositional units in a text must be connected by rhetorical relations in some way for the text to be coherent. By propositional unit, or simply proposition, we mean the meaning of a text segment, usually a clause. The connection of all the text propositions produces its rhetorical/discourse structure.

When there are unconnected parts in a text, it happens that the text presents non-sequitur parts (since it is not possible to relate them), which attribute some level of incoherence to the text. For this reason, it is said that a text must present at least one possible rhetorical structure in order to be classified as coherent.

Rhetorical structures are usually represented by (binary or not) trees, where internal nodes are rhetorical relations and the leaves are propositional units. Although trees are by far the dominant representation in RST works, some researchers argue that graphs should be the most appropriate formalism because they can also represent relationships between subtrees (for details on this subject, see [50]). In DiZer, trees are used to represent rhetorical structures.

As an example of a rhetorical analysis of a text, consider Text 1 in Figure 1 (with numbered segments that express the propositional units) and a possible rhetorical structure in Figure 2.

> [1] Although he is allergic to it, [2] he tried it. [3] Now, he has a headache and [4] his body is red.

Figure 1. Text 1

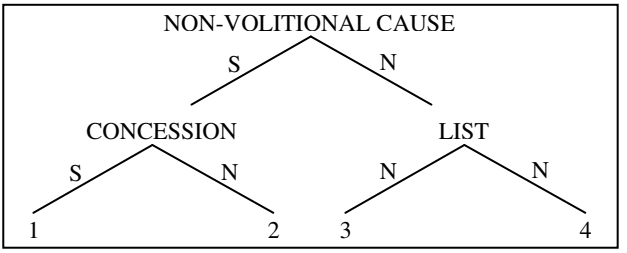

Figure 2. Text 1 rhetorical structure

The symbols N and S indicate the nucleus and satellite of each rhetorical relation: in RST, the nucleus indicates the most important information in the relation, while the satellite provides complementary information to the nucleus. In this structure, propositions 1 and 2 are in a CONCESSION relation, i.e., the fact of being allergic to something should avoid someone of trying it; propositions 1 and 2 CAUSE (not volitionally) propositions 3 and 4; propositions 3 and 4 present a LIST of allergy symptoms. In some cases, relations are multinuclear (e.g., LIST relation), that is, they have no satellites and the connected propositions have the same importance; otherwise, relations are mononuclear, with one nucleus and one satellite (e.g., CONCESSION and NON-VOLITIONAL CAUSE relations). RST originally defines about 25 relations. DiZer uses 32 relations.

One important point about RST that must be mentioned is that, in order to guarantee the construction of valid and well-formed rhetorical structures during the analysis of texts, Mann and Thompson established the compositionality criterion. It says that, for connecting two subtrees T1 and T2 by a relation R in order to form a bigger tree T3, R must hold between the most salient propositional units of T1 and T2, i.e., R must relate the most nuclear units of subtrees T1 and T2. For example, in Figure 2, to form the complete tree, the NON-VOLITIONAL CAUSE relation must hold between the most salient units of the subtrees headed by the CONCESSION and LIST relations, i.e., it must connect units 2 (from the left subtree) and 3 and 4 (from the right subtree). If in the text the NON-VOLITIONAL CAUSE would relate units 1 (which is a satellite from the left subtree and, therefore, is not the most salient unit in the subtree) and units 3 and 4 (from the right subtree), the structure in Figure 2 would be an invalid structure, since it would violate the criterion. This will be further discussed in Section 4.

In terms of knowledge level and other proposals on discourse representation, RST is in the middle in the language production/interpretation process, i.e., rhetoric is the level that bonds the writer intentions to the interpropositional semantics in a text: the intentions specify how a text must be rhetorically organized in order to the reader recognize such intentions and act adequately; rhetorical organization, on its turn, must be established over the semantic organization of a text, that is, it is only possible to rhetorically connect two propositions if they are already semantically connected. It is important to notice the difference between rhetoric and semantics in this view: although both are of interpropositional nature, rhetoric has argumentative force (reflecting the writer intention), while semantics merely expresses relations among factual information. Although it is not a complete consensus, several works

support this functional perspective and propose possible mappings among the referred discourse levels (e.g., [18][22][24][29][30][31][35][42]).

2.2 Discourse parsers: methods and knowledge sources

As previously mentioned, some discourse parsers based on RST are available for both English (e.g., [7][19][21][23][26][41][45][46]) and Japanese languages (e.g., [47]).

Marcu's parser [21][23] was the first representative one and was developed for English news texts. Its development methodology and results were the basis and the motivation for other parsers to arise, since it was showed that good quality discourse parsing is a reachable goal. The parser is a symbolic one, using rules for identifying the discourse relations and the text segments that express the connected propositions. Learned from a corpus analysis, such rules make use of cue phrases in the text, more specifically, the discourse markers, which are, by excellence, the best signals of the discourse structuring of texts. For instance, the markers "but" and "however" indicate CONTRAST relations among the propositions that are expressed by the segments in which they appear, and "therefore" and "as consequence" indicate CAUSE or RESULT relations (volitional or not). When no marker is available among some text segments, Marcu's parser hypothesize the most generic relation in his relation set (which is the ELABORATION relation).

Several works have studied the discourse makers function in discourse (see, e.g., [9][10][12][14][33][40]). The discourse markers are like hints that the writer leaves in the text for the reader to better and faster identify its discourse structure and, consequently, recognize the writer intention and perceive the message the text is supposed to convey.

At this point, it is important to notice that there is not a one-to-one mapping between discourse markers and rhetorical relations: in the same way that CONTRAST relation may be signalled by "but" or "however", these markers can also signal other relations like CONCESSON, ANTITHESIS and OTHERWISE. This causes the fact that discourse markers-based parsers, as Marcu's parser, allow for the multiplicity of discourse analysis for the same text. To reduce the number of analysis, Marcu's parser uses the compositionality criterion, avoiding, this way, not perfectly valid rhetorical structure to be produced. In fact, this criterion is embedded in an intermediary algorithm that the parser uses to build the complete rhetorical structures from the individual relations among propositions previously identified by the rules. From these individual relations, this algorithm produces a logic grammar that generates all the valid structures.

Marcu's parser was followed by Microsoft® initiative [7], which gave rise to RASTA (Rhetorical Structure Theory Analyzer), the first system commercially used. Based on Marcu's parsing method, it also uses aspects of sentences syntactic analysis and their logical form. For instance, to establish a CAUSE relation among two propositions, besides the presence of the phrases "cause" or "results from", the corresponding segments must not be syntactically subordinated to each other and one of the segments must be in the passive voice. RASTA was developed for encyclopaedic texts.

Some discourse parsers employ machine learning techniques. Marcu and Echihabi [26], Soricut and Marcu [46] and Reitter [41] systems are examples of this approach. Worthy of special attention is the work of Soricut and Marcu, which presents a statistical model for

intra-sentential discourse parsing that, with manually pre-edited data, achieves near human-level performance for English news texts, using syntactic and lexical features. However, this model seems not to be extensible to inter-sentential analysis.

The system we present in this paper, DiZer, may be classified as a cue phrase-based parser. Compared to the above works, DiZer applies a differentiated analysis method and uses other knowledge sources besides the discourse markers. The system is described in what follows.

3 DiZer

The main knowledge source in DiZer is a rhetorical repository, which comprises all the knowledge the parser uses. It is codified in form of discourse templates, heuristics and word classes, which were manually produced from a corpus analysis. Using such repository, DiZer applies an incremental analysis strategy to produce the possible discourse structures of texts. In the following subsection, the corpus annotation and the knowledge extraction process to build the rhetorical repository are described. In subsection 3.2, we present DiZer in details and show how the rhetorical repository is used and how the automatic analysis is performed.

3.1 Rhetorical repository

The rhetorical repository codifies the correspondence between discourse structuring and text features. For building this repository, a rhetorically annotated corpus was manually analyzed.

The selected corpus is composed of 100 scientific texts on Computer Science taken from monograph introductory sections (c.a. 53.000 words and 1.350 sentences). The scientific genre was chosen for the following reasons: a) scientific texts are supposedly well written; b) they usually present more cue phrases than other text genres; c) other projects on discourse for Brazilian Portuguese (e.g., [11][38]) use the same sort of texts.

The corpus was rhetorically annotated following Carlson and Marcu's discourse annotation manual [6], which consists of a collection of annotation rules identified by experts who worked on the development of Marcu's parser [21][23]. Although this manual focuses on the English language, it may also be applied to other languages, since RST is theoretically language independent. The use of this manual has allowed a more systematic and mistake-free annotation.

For annotating the texts, Marcu's adaptation of RSTTool [32] was used as the support edition tool. To guarantee consistency during the annotation process, the corpus was annotated by only one expert in RST. A possible problem with this approach is that, in the system evaluation, the use of a portion of this same corpus would introduce some bias in the results. To avoid this, we use another corpus for evaluation, which is a reference corpus, as will be discussed later. This annotation and evaluation strategy is also followed in other works on discourse (see, e.g., [48]).

Initially, for annotating the corpus, the original RST relation set was used. When necessary, more relations were added to the set. In the end, the full set amounted to 32 relations, as shown in Table 1. The added ones are in italics. Some of them (PARENTHETICAL and SAME-UNIT) are only used for organizing the discourse structure. The table also shows the frequency of each relation in the analyzed corpus. One may see that ELABORATION and LIST are the most common relations, as also happens in related works. Interestingly, the JOINT relation does not appear in the corpus, while others have very low frequency, as OTHERWISE and SUMMARY.

Table1. DiZer relation set

Relation	Frequency (%)
ANTITHESIS	0.43
ATTRIBUTION	3.81
BACKGROUND	2.33
CIRCUMSTANCE	3.13
COMPARISON	0.23
CONCESSION	1.46
CONCLUSION	0.29
CONDITION	0.41
CONTRAST	1.83
ELABORATION	34.64
ENABLEMENT	1.09
EVALUATION	0.31
EVIDENCE	0.31
EXPLANATION	0.62
INTERPRETATION	0.29
JOINT	0
JUSTIFY	1.98
LIST	11.33
MEANS	1.36
MOTIVATION	0.39
NON-VOLITIONAL CAUSE	1.36
NON-VOLTIONAL RESULT	0.78
OTHERWISE	0.04
PARENTHETICAL	7.42
PURPOSE	9.42
RESTATEMENT	0.41
SAME-UNIT	8.10
SEQUENCE	1.44
SOLUTIONHOOD	1.03
SUMMARY	0.08
VOLITIONAL CAUSE	1.71
VOLITIONAL RESULT	1.96

The annotation strategy for each text was incremental, in the following way: initially, all propositions of each sentence were related by rhetorical relations; then, the sentences of each paragraph were related; finally, the paragraphs were related. This annotation scheme takes advantage of the fact that the writer tends to put together (i.e., in the same level in the hierarchical organization of the text) the related propositions. For instance, if two propositions are directly related (e.g., a cause and its consequence), it is probable that they will be expressed in the same sentence or in adjacent sentences. This same reasoning is used in DiZer for analyzing texts. For both corpus annotation and automatic analysis, this strategy imposes a restriction on the rhetorical relationships that RST allows, but significantly reduces the number of possible rhetorical structures, which is desired in parsing tasks.

Once completely annotated, the corpus was manually analyzed in order to identify cue phrases, i.e., discourse markers and indicative phrases and words, and heuristics that might indicate rhetorical relations. At this point, it is important to notice that in this work indicative phrases and words play a distinct role. While discourse markers directly signal the discourse structure without affecting the semantics of the proposition they belong to, indicative phrases and words are groups of words that indicate what the expected meaning of the text segment they belong to is [34], i.e., they are part of the propositional content. Table 2 shows the percentage of relations in the corpus that are superficially marked with cue phrases. Notice that volitional and non-volitional CAUSE and RESULT relations are unified in these counts.

Table 2. Percentage of marked relations in the corpus

Relation	Number of marked relations	% of marked relations
ANTITHESIS	20	95,2
ATTRIBUTION	185	100
BACKGROUND	47	41,5
CAUSE	147	98,6
CIRCUMSTANCE	138	90,0
COMPARISON	11	100
CONCESSION	67	94,3
CONCLUSION	12	85,7
CONDITION	20	100
ELABORATION	1.010	60,0
ENABLEMENT	47	88,6
EVALUATION	14	93,3
EVIDENCE	3	20,0
EXPLANATION	23	76,6
INTERPRETATION	12	85,7
JUSTIFY	91	94,7
MEANS	60	90,9
MOTIVATION	16	84,2
OTHERWISE	2	100
PURPOSE	450	98,4
RESTATEMENT	17	85,0
RESULT	129	96,9

SOLUTIONHOOD	49	98,0
SUMMARY	4	100
CONTRAST	83	93,2
LIST	256	46,5
SEQUENCE	51	72,8

Table 3 shows the distribution of cue phrases in the nuclei and satellites of the mononuclear relations. Table 4 shows the same for the multinuclear relations. It is interesting to notice that for some relations the cue phrases appear only on the nucleus or only on the satellite. For instance, cue phrases of ATTRIBUTION relations only appear on the satellites, while in 95.9% of the SOLUTIONHOOD relations, both nucleus and satellite present cue phrases.

Table 3. Percentage of nuclei and satellites in mononuclear relations with cue phrases

Relation	% of nuclei with cue phrase	% of satellites with cue phrase	% of nuclei and satellites with cue phrases
ANTITHESIS	85,0	15,0	0
ATTRIBUTION	0	100	0
BACKGROUND	76,6	8,5	14,9
CAUSE	45,6	24,4	30,0
CIRCUMSTANCE	11,6	80,4	8,0
COMPARISON	0	45,4	54,6
CONCESSION	35,8	56,7	7,5
CONCLUSION	0	100	0
CONDITION	0	90,0	10,0
ELABORATION	0	99,3	0,7
ENABLEMENT	83,0	14,9	2,1
EVALUATION	0	100	0
EVIDENCE	0	100	0
EXPLANATION	0	100	0
INTERPRETATION	0	100	0
JUSTIFY	8,8	9,9	81,3
MEANS	1,7	98,3	0
MOTIVATION	81,2	18,8	0
OTHERWISE	0	100	0
PURPOSE	0	97,3	2,7
RESTATEMENT	5,9	94,1	0
RESULT	3,9	93,8	2,3
SOLUTIONHOOD	0	4,1	95,9
SUMMARY	0	100	0

Table 4. Percentage of nuclei in multinuclear relations with cue phrases

Relation	% of 1st nuclei with cue phrases	% of 2nd nuclei with cue phrases	% of both nuclei with cue phrases
CONTRAST	1,2	97,6	1,2
LIST	0,8	80,5	18,7
SEQUENCE	2,0	88,2	9,8

Table 5 shows the percentage of relations for which the nucleus appears before the corresponding satellite in the text and vice-versa, i.e., the preferential order of nuclei and satellites of the relations. It is also interesting to see that some relations always have their nucleus first and other always have their satellite first in the text. For example, the CONCLUSION relation never presents the satellite first, what corresponds to our intuition, since a conclusion is always presented after some preliminary arguments.

Table 5. Preferential order of nuclei and satellites of the relations

Relation	Nucleus before satellite (%)	Satellite before nucleus (%)
ANTITHESIS	14,2	85,8
ATTRIBUTION	2,7	97,3
BACKGROUND	0,9	99,1
CAUSE	24,8	75,2
CIRCUMSTANCE	49,3	50,7
COMPARISON	91,0	9,1
CONCESSION	19,7	80,3
CONCLUSION	100	0
CONDITION	50,0	50,0
ELABORATION	99,7	0,3
ENABLEMENT	24,5	75,5
EVALUATION	100	0
EVIDENCE	100	0
EXPLANATION	100	0
INTERPRETATION	100	0
JUSTIFY	78,1	21,9
MEANS	86,4	13,6
MOTIVATION	21,,0	79,0
OTHERWISE	100	0
PURPOSE	85,3	14,7
RESTATEMENT	95,0	5,0
RESULT	96,2	3,8
SOLUTIONHOOD	0	100
SUMMARY	100	0

The cue phrases identified in the corpus analysis yielded the discourse templates that DiZer uses, amounting to about 750 templates. Moreover, some heuristics were designed for some relations that are usually not superficially signalled in texts.

In Figure 3, we show an example of a template for the OTHERWISE rhetorical relation. According to it, an OTHERWISE relation connects two propositional units 1 and 2, with 1 being the satellite and 2 the nucleus and with the segment that expresses 1 appearing before the segment that expresses 2 in the text, if the discourse marker *ou, alternativamente* ("or, alternatively", in English) appears in the beginning of the segment that expresses propositional unit 2.

Relation	OTHERWISE
Order	satellite (S) before nucleus (N)
1st marker	---
Position of 1st marker	---
2nd marker	*ou, alternativamente*
Position of 2nd marker	beginning

Figure 3. Discourse template for the OTHERWISE relation

Such template came from text spans like the one below (translated from Portuguese[2]), with the discourse marker in bold:

> "To produce a summary, the relevant information in the text must be identified in order to compose the summary, **or, alternatively**, irrelevant information in the text must be identified and omitted in the summary."

The idea is that, when a new text is given as input to DiZer, a pattern matching process is carried out. If one of the templates matches some portion of the text being processed, the corresponding rhetorical relation is supposed to occur between the appropriate segments.

The templates may also convey morphosyntactic information, lemma and specific genre-related information. For instance, consider the template in Figure 4, which hypothesizes a PURPOSE relation.

Relation	PURPOSE
Order	satellite (S) before nucleus (N)
1st marker	---
Position of 1st marker	---
2nd marker	lem(*cujo*) PurposeWord * ADJ lem(*ser*)
Position of 2nd marker	beginning

Figure 4. Discourse template for the PURPOSE relation

This template specifies that a PURPOSE rhetorical relation is found if there is in the text an indicative phrase composed by (1) a word whose lemma is *cujo* ("whose" or "which", in English), (2) followed by any word that indicates purpose, which is represented by the

[2] The original version in Portuguese is: *Assim, para produzir sumários deve-se identificar, no texto-fonte, as informações mais relevantes que devem compor o sumário ou, alternativamente, identificar as informações menos relevantes que devem ser omitidas no sumário.*

PurposeWord word class, which includes, for instance, words like *propósito* ("purpose", in English), *objetivo* ("objective" or "goal") and *proposta* ("proposal"), (3) followed by any number of words, which is indicated by the mask character *, (4) followed by any adjective, (5) followed by a word whose lemma is *ser* (verb "to be", in English). In English, this would correspond to an indicative phrase such "whose main goal is" (notice that, in English, the adjective comes before the noun, while in Portuguese the opposite generally occurs). An example of text span that represents this template is (translated from Portuguese[3]):

> "This dissertation is the result of a work **whose main goal is** to investigate the application of constructive neural networks to pattern recognition tasks."

The use of word classes in the templates (instead defining whole lexicalized templates) allows the definition of fewer and more general templates. In DiZer, there are word classes that indicate, for instance, research related words (e.g., in English, "research", "study", "survey" and "investigation"), causative verbs (e.g., "to cause", "to make", "to result" and "to provoke"), evidence verbs (e.g., "to evidence", "to demonstrate" and "to attest"), among others. A complete list of these word classes may be found in [36].

The mask character attributes some flexibility to the templates, allowing the occurrence of long distance dependencies in indicative phrases. In the example, the mask character would allow the occurrence of other words between PurposeWord and the adjective. This, in English, would result in phrases such as "whose main expected goal is", in which the word "expected" matches the mask character.

For the EVALUATION and SOLUTIONHOOD relations, which generally do not have corresponding cue phrases, it was possible to define some heuristics to enable the discourse parsing, given the specific text genre under focus. For the SOLUTIONHOOD relation, for example, the following heuristic (adapted to English) holds:

> if in a segment X, "negative" words like "cost" and "problem" appear more than once and, in segment Y, which follows X, "positive" words like "solution" and "development" appear more than once too, then a SOLUTIONHOOD relation holds between propositions expressed by segments X and Y, with X being the satellite and Y the nucleus of the relation

Another example is the heuristics for EVALUATION below:

> if in a segment X, "evaluative" words like "satisfactory", "adequate" and "success" appear more than once, then a EVALUATION relation holds between propositions expressed by segments X and Y, with X following Y in the text and with X being the satellite of the relation

[3] The original version in Portuguese is: *Esta dissertação é o resultado de um trabalho cujo objetivo inicial é investigar a aplicação de Redes Neurais Construtivas, RNCs, em tarefas de Reconhecimento de Padrões.*

The following text span is a representative example of this heuristic:

> "The system based on conventional radio control technology was developed for agriculture. The experiment we carried out indicates that the system is **satisfactory**, with **success** in several tasks."

The second sentence is in an EVALUATION relation with the first sentence. The evaluative words are in bold.

One can see that, in fact, what the above heuristics do is to look for indicative words in the segments.

Next section describes DiZer and its processes, showing how and where the rhetorical repository is used.

3.2 DiZer architecture

DiZer comprises three main processes: (1) the segmentation of the text into propositional units, (2) the detection of rhetorical relations between propositional units and (3) the building of the rhetorical structures. Figure 5 presents the system architecture. Following it, a source text to be parsed is first POS tagged and segmented into text spans, which express the propositions; next, the rhetorical relations between the propositions are detected; finally, the overall structure is built. In the next subsections, each process is explained in detail. The information repositories are introduced as the processes that use them are explained.

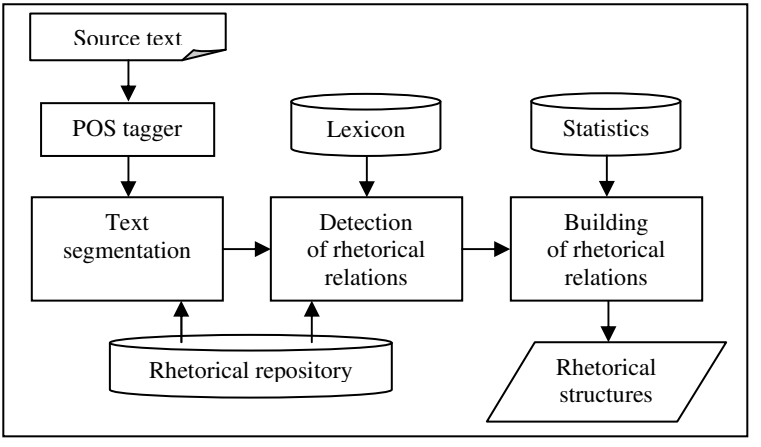

Figure 5. DiZer architecture

3.2.1 *Text segmentation*

In this process, DiZer tries to determine the simple clauses in the source text, since they usually express single propositional units.

DiZer initially assigns morphosyntactic categories to each word in the text using a Brazilian Portuguese part-of-speech tagger [1]. Then, the segmentation process is carried out, segmenting the text always a punctuation signal (comma, dot, exclamation and interrogation points, etc.) or a strong cue phrase is found. Given the ambiguity of dot, an abbreviation list is used to identify which dots are sentence boundaries and not abbreviation terminations. By strong cue phrase, we mean those words that unambiguously play a role in discourse, clearly indicating a rhetorical relation between propositions or signalling the discourse structure. According to this, words like *e* (in English, "and") are ignored, while words like *portanto* and *por exemplo* (in English, "therefore" and "for instance", respectively) are not. The cue phrases are retrieved from the rhetorical repository. DiZer still verifies whether the identified segments are clauses by looking for occurrences of verbs in them.

Optionally, in DiZer, it is also possible to perform sentential segmentation instead of clausal segmentation.

As will be seen in the next section, the results produced by this simple segmentation process are good. However, a syntactic parser could perform better in identifying segments, since it is able to detect clauses. We have not used a parser because, when DiZer was built, there was no free and robust parser available for Portuguese. Such scenario has changed, however, and we plan to integrate a parser in the system in the near future.

3.2.2 *Detection of rhetorical relations*

In order to look for rhetorical relations, DiZer makes use of the discourse templates in the rhetorical repository. It performs a pattern matching process between text segments and the templates. When the templates require lemma information, a Brazilian Portuguese lexicon [27] is consulted. After the pattern matching, if no relation was detected for some segments, heuristics are applied for these cases.

DiZer tries to determine at least one rhetorical relation for each two adjacent text segments representing the corresponding underlying propositions. Initially, in its incremental analysis, it looks for a relation between every two adjacent clauses in a sentence; then, it considers every two adjacent sentences of a paragraph; finally, it considers every two adjacent paragraphs. As already discussed, such strategy limits RST expressive power, but reduces the number of possible analyses for a text.

When more than one rhetorical relation is detected for two segments, usually in occurrences of ambiguous or multiple cue phrases, all the possible relations are considered. Because of this, several discourse structures may be produced for the same text. In the worst case, when no rhetorical relation can be found between two segments, DiZer assumes a default heuristic: it adopts an ELABORATION relation (which is the most generic relation in DiZer relation set), with the first segment being its nucleus.

3.2.3 *Building of rhetorical structures*

This process consists in building the complete rhetorical structure from the individual rhetorical relations between the text segments. For this, the system makes use of the

algorithm proposed by Marcu [21]. This algorithm produces grammar rules for each possible combination of segments by a rhetorical relation, in the form of a DCG (Definite-Clause Grammar) rule [39]. When the grammar is executed, all possible valid rhetorical structures are built.

Marcu's algorithm incorporates the compositionality criterion established by RST. In DiZer, this criterion is ignored when it shows to be too restrictive to allow the production of any rhetorical structure, as will be discussed in the next section.

In the end of this process, DiZer offers the possibility of ranking all the produced structures according to their probabilities. The probability of a structure is simply the product of the probabilities of each relation and its immediate children (with their nuclearity indication) in the tree, which can be other relations or leaves (if they are terminal nodes). These probabilities are stored in the statistical repository and are simple frequency counts collected from the rhetorically annotated corpus from which the discourse templates were extracted. They are conditional probabilities, i.e., they codify the probability of the occurrence of the children and their nuclearity status – nucleus or satellite – given the parent. When a probability is required but is not found in the repository, a very low probability (which was empirically established as 10^{-6}) is used, guaranteeing that rhetorical structures have non-zero probabilities.

Formula 1 defines the probability calculus for a given rhetorical structure T, where LC and RC hold for immediate Left Child and Right Child node labels (for a relation – internal node – in the structure T), and Status indicates whether a child is a "nucleus" (N) or a "satellite" (S):

Formula 1: $$prob(T) = \prod_{\text{for each relation } R \text{ in } T} prob(LC, StatusLC, RC, StatusRC \mid R)$$

LC and RC values may be relation names, if the children are internal nodes in the structure, or "leaf", if they are terminal nodes in the structure.

As a complete example of DiZer processing, Figures 6 and 7 present, respectively, a text (translated from Portuguese[4]) already segmented by DiZer and one of the valid rhetorical structures built. One may verify that the structure is totally plausible.

[4] The original version in Portuguese is: *Desde a sua abertura comercial, em 1993, a Internet tornou-se um meio de comunicação poderoso, ao permitir a um usuário entrar em contato com quaisquer outros, espalhados pelo mundo todo. O comércio eletrônico é um dos novos nichos de exploração comercial da rede mundial de computadores, pois ela torna possível realizar transações comerciais de forma global, com custo de manutenção inferior ao empregado em uma rede de comércio tradicional.*

O objetivo deste trabalho é apresentar uma proposta para o projeto e implementação de um serviço de comércio eletrônico na plataforma JAMP. Esta plataforma constitui-se em um middleware implementado em Java/RMI para desenvolvimento de aplicações multimídia distribuídas, e em particular, aplicações para World Wide Web (WWW), através de frameworks de serviços para suporte ao desenvolvimento destas aplicações.

[1] Since its commercial opening at 1993, Internet became a powerful communication service [2] when permitted a user to get in touch with any other users in the world. [3] The electronic commerce is one of the new exploration niches in Internet, [4] because Internet makes it possible to realize global commercial transactions with inferior maintenance cost.

[5] The purpose of this work is to propose the project and implementation of an electronic commerce service on the JAMP platform. [6] This platform is a middleware implemented on Java/RMI for distributed multimedia applications development and, in particular, for World Wide Web applications, through service frameworks for these applications development support.

Figure 6. Text 2

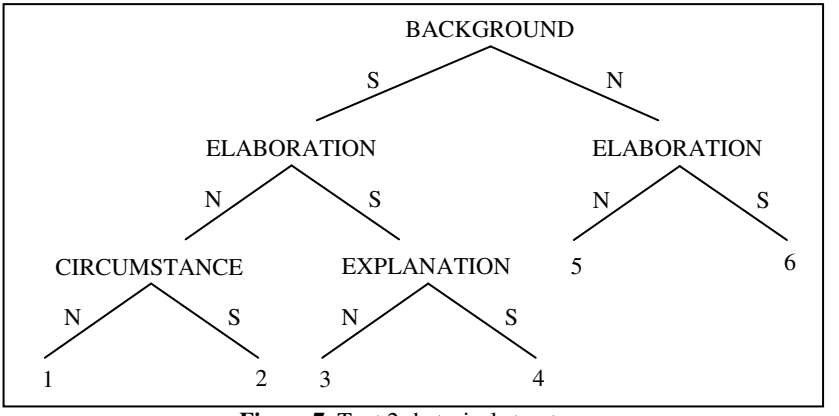

Figure 7. Text 2 rhetorical structure

The probability of such structure would be computed as the following:

P(structure) = P(ELABORATION,S,ELABORATION,N|BACKGROUND) x
P(CIRCUMSTANCE,N,EXPLANATION,S|ELABORATION) x
P(leaf,N,leaf,S|ELABORATION) x
P(leaf,N,leaf,S|CIRCUMSTANCE) x
P(leaf,N,leaf,S|EXPLANATION)

Next section describes DiZer evaluation.

4 DiZer evaluation

In order to evaluate DiZer, a reference corpus was produced. The corpus, called Rhetalho [37], is composed of 50 rhetorically annotated texts (with size over half a page long) from scientific and news genres, which are different from the ones used to develop DiZer rhetorical repository. The scientific texts are also from Computer Science domain; the news texts were collected from several sections of the on-line newspaper *Folha de São Paulo*.

All the texts were annotated by two judges (experts in RST), using Marcu's RSTTool adaptation and following an annotation protocol in order to achieve agreement. The protocol specifies the following:

- the annotation of a text must be linear, from left to right, and incremental;
- whenever possible, as soon as a new segment is determined, it must be related to the tree already built until that point;
- only binary structures are allowed, i.e., each node in the tree may have up to 2 children; when a non-binary tree is produced, it must be transformed into a binary tree (for instance, a CONTRAST relation with 3 children should be transformed in a CONTRAST relation with 2 children, with one being the first child and the other being another CONTRAST relation connecting the 2 remaining children);
- for segmenting a text, the rules defined in [6] must be followed;
- when the judges disagree about a segment, the larger segment must be chosen;
- when judges hypothesize different relations for connecting two segments, the most generic one must be chosen; when they are equally generic and plausible, a third judge must be consulted.

DiZer was evaluated with 20 scientific texts and 5 news texts (from Section "World") randomly selected from Rhetalho. The evaluation with news texts was conducted in order to verify the possibility of using DiZer with other text genres and domains, since it was developed based only on a corpus of Computer Science scientific texts.

The traditional measures recall and precision were computed for the main aspects of the rhetorical structures produced by DiZer, namely, delimited segments, nuclearity of segments and detected rhetorical relations. This was done for both clausal and sentential segmentation in DiZer.

For text segmentation, recall indicates how many segments of the reference structure (from Rhetalho) were correctly delimited and precision indicates how many of the delimited segments were correct; for nuclearity of segments, recall indicates how many nuclei and satellites of the reference structure were correctly identified and precision indicates how many of the segments were correctly classified (as nuclei or satellites); for rhetorical relations detection, recall indicates how many relations between segments of the reference structure were correctly detected and precision indicates how many of the detected relations were correct.

In order to measure the validity of DiZer results, we ran the same evaluation for a standard baseline method, which performs sentential segmentation and detects only ELABORATION relations, with the first segment being the nucleus.

Table 6 presents the resulting recall (R) and precision (P) average numbers for the baseline method and for DiZer analyses with clausal and sentential segmentation for scientific texts. Table 7 presents the figures for the news texts. It also includes f-measure (F), which is a combination of recall and precision and can be an indication of how good a system is.

Table 6. DiZer performance for scientific texts

Evaluated aspect	DiZer with sentential segmentation (%)			DiZer with clausal segmentation (%)			Baseline method (%)		
	R	P	F	R	P	F	R	P	F
Segmentation	25.2	41.7	31.4	57.3	56.2	56.8	25.2	41.7	31.4
Nuclearity	39.1	69.5	50.1	79.7	82.3	80.9	32.4	59.5	42.0
Relations	28.7	61.0	39.1	63.2	61.9	62.5	20.7	49.2	29.2

Table 7. DiZer performance for news texts

Evaluated aspect	DiZer with sentential segmentation (%)			DiZer with clausal segmentation (%)			Baseline method (%)		
	R	P	F	R	P	F	R	P	F
Segmentation	9.9	20.6	13.4	48.8	54.1	51.3	9.9	20.6	13.4
Nuclearity	22.3	55.3	31.8	55.8	63.5	59.4	28.4	71.3	40.7
Relations	12.5	38.3	18.9	37.8	43.2	40.3	17.6	58.3	27.0

According to the f-measure values, for scientific texts, DiZer outperformed the baseline method for both sentential and clausal segmentation, with very good results for the latter. For the news texts, DiZer outperformed the baseline method for the clausal segmentation only. We believe that these bad results for sentential segmentation are due to the way news texts are organized: most of the relations in news texts are ELABORATION, with the first segment being the nucleus, which is exactly the way the baseline method works.

In general, the clausal segmentation outperforms the sentential segmentation because it enables DiZer to produce more fine-grained structures, which are closer to Rhetalho reference structures.

DiZer performance shows to be satisfactory even for news texts, when clausal segmentation is carried out, overcoming the baseline method. It also conforms to other literature results, especially to Marcu's parser [21][23], which is the most similar to DiZer in literature. Table 8 shows the overall results reported by the cited related work in Section 2. Although such direct comparisons are unfair, given that languages and test corpora differ, it gives an idea of the state of the art results in cue phrase-based parsers.

Table 8. Literature results

Evaluated aspect	Performance
Segmentation	84-97%
Nuclearity	63%
Relations	49-75%

It is possible to see that DiZer segmentation performance (56.8% for scientific texts with clausal segmentation) is far below the results achieved in the area. The use of a syntactic parser can certainly solve this. In terms of nuclearity (the system achieved 80.9% for scientific texts with clausal segmentation), DiZer is above the reported results. Concerning the relations detection (the system achieved 62.5% for scientific texts with clausal segmentation), the system is within the obtained interval.

We have identified some causes for DiZer parsing errors. In clausal segmentation, the lack of a syntactic parser does not allow the exact determination of clause boundaries; simple rules based on punctuation signals are not enough for achieving very good results. In rhetorical relations detection, most of segments do not contain cue phrases, which causes the generation of a big amount of ELABORATION relations. Still, if the tagger fails in identifying the morphosyntactic classes of words (its precision is about 89%), discourse parsing may be affected during clausal segmentation (if verbs are not correctly classified) and rhetorical relations detection (when a discourse template asks for morphosyntactic classes that may be wrong in the sentence). Another problem, not so frequent in our test corpus, is related to the quality of the text to be parsed: in some cases, cue phrases are misused, which introduces errors during rhetorical relations detection. This is specially true for the scientific texts, which are written by graduate students. On the other hand, news texts are written by professional writers, which is not exactly the case of Computer Scientists in general.

During DiZer evaluation, we also verified how many times the compositionality criterion could be applied. For scientific texts, the criterion was applied in 75% of the cases for sentential segmentation and in only 20% of the cases for clausal segmentation; for news texts, the criterion was applied in 60% of the cases for sentential segmentation and in only 20% of the cases for clausal segmentation. If DiZer were unable to ignore the compositionality criteria when this was too restrictive to allow the production of any rhetorical structure, just a few texts would have their structures produced. In general, we found that the compositionality criterion is desired in theory, but, in an automatic analyzer, it may not be: a single relation or nuclearity that is wrongly hypothesized for a text (which is not rare in automatic discourse parsing, given the subjectivity of texts) may avoid the construction of any structure. A previous work [35] shows that it is possible to have plausible rhetorical structures even when the compositionality criterion is not applied.

5 Final remarks

This paper presented DiZer main aspects and a comprehensive evaluation of the system, which showed satisfactory results. To our knowledge, DiZer is the first discourse parser for Brazilian Portuguese.

Although DiZer was developed for parsing scientific texts, its evaluation shows that it is possible to achieve acceptable results for other text genres and domains. We believe that this happens because cue phrases are consistently used across text genres and domains.

We are now investigating the reproduction of DiZer development methodology to generate a customized version of the system to news texts and preparing the system to have a syntactic parser plugged to it. In terms of applications, DiZer has been applied to anaphora resolution for Brazilian Portuguese [5] and we also aim at using it for text summarization in the near future.

DiZer is the first step towards the automation of other levels of discourse parsing. As suggested in [35], it is possible to map directly rhetorical relations to the semantic relations proposed in [17]. This should be investigated in the future. Other methods for discourse

parsing should be studied too, for instance, statistical models and machine learning techniques. These methods could also benefit from corpus automatically annotated by DiZer (and manually revised, if necessary).

Acknowledgments

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

References

[1] Aires, R.V.X.; Aluísio, S.M.; Kuhn, D.C.S.; Andreeta, M.L.B.; Oliveira Jr., O.N. (2000). Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In the *Proceedings of the Brazilian AI Symposium* – SBIA, pp. 20-22.

[2] Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Aarhus: Aarhus University Press.

[3] Bosma, W.E. (2005). Extending Answers using Discourse Structure. In the *Proceedings of the Workshop on Crossing Barriers in Text Summarization Research*, pp. 2-9. Bulgaria.

[4] Burstein, J.; Marcu, D.; Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, pp. 32-39.

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

[6] Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.

[7] Corston-Oliver, S. (1998). *Computing Representations of the Structure of Written Discourse*. PhD Thesis, University of California, Santa Barbara, CA, USA.

[8] Cristea, D.; Ide, N.; Romary, L. (1998): Veins Theory. An Approach to Global Cohesion and Coherence. In the *Proceedings of Coling/ACL.*

[9] Di Eugenio, B. (1992). Understanding natural language instructions: the case of purpose clauses. In the *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics* – ACL, pp. 120-127. Newmark, DE.

[10] Di Eugenio, B. (1993). *Understanding natural language instructions: a computational approach to purpose clauses*. Ph.D. thesis, University of Pennsylvania.

[11] Feltrim, V.D.; Teufel, S.; Nunes, M.G.V.; Aluísio, S.M. (2005). Argumentative Zoning Applied to Critiquing Novices' Scientific Abstracts. In J.G. Shanahan; Y. Qu; J. Wiebe. (Org.), *Computing Attitude and Affect in Text: Theory and Applications*. Berlim: Kluwer, V. 1, pp. 233-244.

[12] Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, V. 32, pp. 913-952.

[13] Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, V. 12, N. 3.

[14] Hirschberg, J. and Litman, D. J. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, V. 19, N. 3, pp. 501-513.

[15] Hovy, E. (1988). *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

[16] Jordan, M.P. (1992). An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W.C. Mann and S.A. Thompson (eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

[17] Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. CSLI Publications.

[18] Korelsky, T. and Kittredge, R. (1993). Towards stratification of RST. In the *Proceedings of the Intentionality and Structure in Discourse Relations Workshop*, pp. 52-55. Ohio, USA.

[19] Mahmud, R. and Ramsay, A. (2005). Finding Discourse Relations in Student Essays. In the *Proceedings of the 6th Computational Linguistics and Intelligent Text Processing International Conference*.

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

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

[22] Marcu, D. (1999). A formal and computational synthesis of Grosz and Sidner's and Mann and Thompson's theories. In the *Proceedings of the Workshop on Levels of Representation in Discourse*, pp. 101-108. Edinburgh, Scotland.

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

[24] Marcu, D. (2000b). Extending a Formal and Computational Model of Rhetorical Structure Theory with Intentional Structures à la Grosz and Sidner. In the *Proceedings of the 18th International Conference on Computational Linguistics* – COLING. Saarbrueken.

[25] Marcu, D.; Carlson, L.; Watanabe, M. (2000). The Automatic Translation of Discourse Structures. In the *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. Seattle, Washington.

[26] Marcu, D. and Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* – ACL, Philadelphia, PA.

[27] Martins, R.T.; Hasegawa, R.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N. (1998). Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. *Natural Language Engineering*, V. 4, pp. 287-307. Cambridge University Press.

[28] Moore, J.D. (1995). *Participating in Explanatory Dialogs: Interpreting and Responding to Questions in Context*. The MIT Press. Cambridge, Massachusetts.

[29] Moore, J.D. and Paris, C. (1993). Plannig Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, V. 19, N. 4, pp. 651-694.

[30] Moore, J. D. and Pollack, M. E. (1992). A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics*, V. 18, N. 4, pp. 537-544.

[31] Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, V. 22, N. 3, pp. 409-419.

[32] 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, Duisburg, Germany.

[33] Oates, S.L. (1999). *State of the Art Reporto n Discourse Markers and Relations*. Technical Report ITRI-99-08. Information Technology Research Institute. University of Brighton.

[34] Paice, C.D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co. (Publishers).

[35] 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, June, 211p.

[36] 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, April, 73p.

[37] Pardo, T.A.S. and Seno, E.M.R. (2005). Rhetalho: um corpus de referência anotado retoricamente. In *Anais do V Encontro de Corpora*. São Carlos-SP, November 24-25.

[38] Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

[39] Pereira, F.C.N. and Warren, D.H.D. (1980). Definite Clause Grammars for Language Analysis – A Survey of the Formalism and Comparison with Augmented Transition Networks. *Artificial Intelligence*, N. 13, pp. 231-278.

[40] Quirk, R.; Greenbaum, S.; Leech, G.; Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, Harcourt.

[41] Reitter, D. (2003). Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *GLDV-Journal for Computational Linguistics and Language Technology*, V. 18, pp. 38-52.

[42] Rino, L.H.M. (1996). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. PhD Thesis. Instituto de Física de São Carlos, Universidade de São Paulo. São Carlos - SP.

[43] Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. (2004). A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In the *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence* – SBIA (Lecture Notes in Artificial Intelligence 3171), pp. 235-244. São Luis-MA, Brazil.

[44] Schauer, H. (2000). Referential Structure and Coherence Structure. In the *Proceedings of TALN*. Lausanne, Switzerland.

[45] Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), *Natural Language Engineering*, V. 8. Cambridge University Press.

[46] Soricut, R. and Marcu, D. (2003). Sentence Level Discourse Parsing using Syntactic and Lexical Information. In the *Proceedings of HLT/NAACL*.

[47] Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S. (1992). A discourse structure analyzer for Japonese text. In the *Proceedings of the International Conference on Fifth Generation Computer Systems*, V. 2, pp. 1133-1140. Tokyo, Japan.

[48] Williams, S. and Reiter, E. (2003). A corpus analysis of discourse relations for Natural Language Generation. In the *Proceedings of Corpus Linguistics*, pp. 899-908. Lancaster University.

[49] Wing, B. and Baldridge, J. (2006). Adaptation of Data and Models for Probabilistic Parsing of Portuguese. In the *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese* – PROPOR, pp. 140-149. Itatiaia, Brazil.

[50] Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based analysis. *Computational Linguistics*, V. 31, N. 2.