Review and Evaluation of DiZer – an Automatic Discourse Analyzer for Brazilian Portuguese

Thiago Alexandre Salgueiro Pardo and Maria das Graças Volpe Nunes

Núcleo Interinstitucional de Lingüística Computacional (NILC)
CP 668 – ICMC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.nilc.icmc.usp.br
taspardo@gmail.com, gracan@icmc.usp.br

Abstract. This paper presents the review and evaluation of DiZer – an automatic discourse analyzer for Brazilian Portuguese. Based on Rhetorical Structure Theory, DiZer is a symbolic analyzer that makes use of linguistic patterns learned from a corpus of scientific texts to identify and build the discourse structure of texts. DiZer evaluation shows satisfactory results for scientific texts. In order to test its portability, DiZer is also evaluated with news texts and presents acceptable performance.

1 Introduction

Researches in Linguistics and Computational Linguistics have shown 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 information conveyed by the text propositional units (i.e., the meaning of the text segments) correlate and make sense together.

The ability to derive discourse structures of texts automatically is of great importance to many applications in Computational Linguistics. For instance, it may be very useful for automatic text summarization (to identify the most important information of a text to produce its summary) (see, for instance, O'Donnel, 1997; Marcu, 2000), co-reference resolution (determining the context of reference in the discourse may help determining the referred term) (see, for instance, Cristea et al., 1998; Schauer, 2000), and for other natural language applications as well.

Some discourse analyzers are already available for both English (e.g., Marcu, 1997, 2000; Corston-Oliver, 1998; Soricut and Marcu, 2003) and Japanese languages, (e.g., Sumita et al., 1992). For English, Marcu's analyzer was the first one available and was developed for free domain texts (based on news texts). To our knowledge, for Brazilian Portuguese, DiZer (DIscourse analyZER) (Pardo et al., 2004) is the only automatic analyzer for this language. Based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), DiZer is a symbolic analyzer that makes use of linguistic patterns learned from a corpus of scientific texts from Computer Science domain to identify and build the discourse structure of texts. Basically, DiZer follows the analysis strategy proposed by Marcu (1997, 2000), using cue-phrases occurrences in a text to build its discourse structure.

In this paper, we review DiZer main aspects and present a comprehensive evaluation of the system. We describe the construction of a reference rhetorically annotated corpus, called Rhetalho (Pardo and Seno, 2005), and the annotation

protocol followed by human judges in order to achieve agreement. DiZer evaluation based on Rhetalho is presented and discussed for both scientific and news texts. Results show that DiZer performance is satisfactory.

Firstly, in the next section, we introduce RST, the discourse theory that DiZer follows. In Section 3, DiZer main processes and information repositories are reviewed. Section 4 describes DiZer evaluation procedure and results. Some conclusions and final remarks are made in Section 5.

2 Rhetorical Structure Theory

There are several discourse theories that try to represent different aspects of discourse (see, e.g., Grosz and Sidner, 1986; Mann and Thompson, 1987; Jordan, 1992; Kehler, 2002). Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) is one of the most used theories and underlies most of the existent automatic discourse analyzers.

According to RST, all propositional units in a text must be connected by rhetorical relations in some way for the text to be coherent. The connection of all the text propositional units produces its rhetorical/discourse structure. Rhetorical structures are usually represented by trees (not necessarily binary), with each relation connecting subtrees, which can be single propositional units (that are leaves in the tree) or other trees. As an example of a rhetorical analysis of a text, consider Text 1 in Figure 1 (with segments that express basic propositional units numbered) and its 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

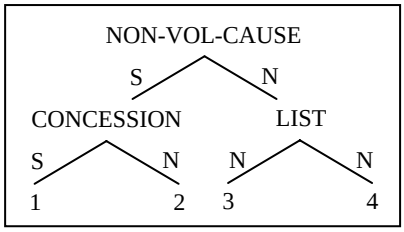

Figure 2. Text 1 rhetorical structure

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-VOL-CAUSE relations). RST originally defines around 25 relations.

One last 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-VOL-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 NON-VOL-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.

As in other automatic discourse analysis works, RST is the discourse theory followed by DiZer, which is reviewed in the next section.

3 DiZer

DiZer comprises three main processes: (1) the segmentation of the text into propositional units, (2) the detection of occurrences of rhetorical relations between propositional units and (3) the building of the rhetorical structures. Figure 3 presents the system architecture. In the next subsections, each process is explained. The information repositories are introduced as the processes that use them are explained.

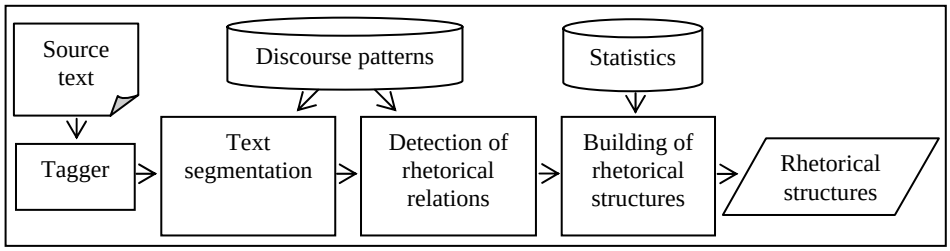

Figure 3. DiZer architecture

3.1 Text Segmentation

In this process, DiZer tries to determine the simple clauses in the source text, since simple clauses usually express single propositional units, which are assumed to be the minimal units in a rhetorical structure. For doing this, DiZer initially attributes morphosyntactic categories to each word in the text using a Brazilian Portuguese tagger (Aires et al., 2000). 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. By strong cue phrase, we mean those words that unambiguously have a function in discourse, clearly indicating a rhetorical relation between propositions or signaling the discourse structure. According to this, words like *e* and *se* (in English, “and” and “if”, respectively) 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 “Discourse patterns” repository, which is better explained in the next subsection. 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 sentence segmentation instead of clause segmentation.

3.2 Detection of Rhetorical Relations

DiZer tries to determine at least one rhetorical relation for each two adjacent text segments representing the corresponding underlying propositions. Initially, it looks for a relation between every two adjacent clauses inside a sentence; then, it considers every two adjacent sentences of a paragraph; finally, it considers every two adjacent paragraphs. This processing order is supported by the premise that a writer organizes related information at the same organization level. 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.

In order to look for rhetorical relations, DiZer makes use of linguistic patterns stored in the “Discourse patterns” repository. Each pattern codifies the possible rhetorical relations that cue phrases may indicate. As an example, consider the discourse pattern for the OTHERWISE rhetorical relation in Figure 4. According to it, an OTHERWISE relation connects two segments 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 cue phrase *ou, alternativamente,* (in English, “or, alternatively,”) be present in the beginning of the segment that expresses propositional unit 2.

Relation	OTHERWISE
Order	satellite (S) before nucleus (N)
Marker1	---
Position of marker1	---
Marker2	*ou, alternativamente,*
Position of marker2	beginning

Figure 4. Discourse pattern for the OTHERWISE rhetorical relation

The discourse patterns may also convey morphosyntactic information, lemma and specific genre-related information. For instance, consider the pattern in Figure 5, which hypothesizes a PURPOSE relation. This pattern specifies that a PURPOSE rhetorical relation is found if there is in the text an cue phrase composed by (1) a word whose lemma is *cujo* (“which”, in English), (2) followed by any word that indicates purpose (represented by the “purWord” class), (3) followed by any adjective, (4) followed by a word whose lemma is *ser* (verb “to be”, in English).

Relation	PURPOSE
Order	satellite (S) before nucleus (N)
Marker1	---
Position of marker1	---
Marker2	*cujo_lem purWord _adj ser_lem*
Position of marker2	beginning

Figure 5. Discourse analysis pattern for the PURPOSE rhetorical relation

For detecting the relations, DiZer performs a pattern matching process between text segments and the discourse patters.

For relations that are not explicitly signaled by cue phrases, like EVALUATION and SOLUTIONHOOD, DiZer uses heuristics. For the SOLUTIONHOOD relation, for example, the following heuristic 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

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, with the segment that appears first in the text being its nucleus.

The discourse patterns and heuristics were produced by manually annotating and analyzing a corpus of 100 Computer Science scientific texts in Brazilian Portuguese, called CorpusTCC (Pardo and Nunes, 2003, 2004). More details about this corpus and the knowledge extraction process to produce the patterns and heuristics can be found in Pardo et al. (2004).

3.3 Building of Rhetorical Structures

This process consists in determining the complete rhetorical structure from the individual rhetorical relations between the text segments. For this, the system makes use of the rule-based algorithm proposed by Marcu (1997). 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 (Pereira and Warren, 1980). When the grammar is executed, all possible valid rhetorical structures are built.

Marcu's algorithm incorporates the compositionality criterion established by RST (see Section 2). In DiZer, this criterion is ignored when it shows to be to 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 by their probabilities. The probability of a structure is simply the multiplication of the probabilities of each relation and their immediate children (with their nuclearity indication) in the tree, which can be other relations or leaves (if they are terminal nodes). These probabilities are simple frequency counts collected from CorpusTCC and are stored in the "Statistics" repository in the form of conditional probabilities (i.e., the probability of the children and their nuclearity given the parent). When a probability is required and is not found in the repository, a very low probability (which was empirically established as 10^{-6}) is used, guaranteeing that the rhetorical structure have a non-zero probability.

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

[*Desde a sua abertura comercial, em 1993, a Internet tornou-se um meio de comunicação poderoso,*]$_1$ [*ao permitir a um usuário entrar em contato com quaisquer outros, espalhados pelo mundo todo.*]$_2$

[*O comércio eletrônico é um dos novos nichos de exploração comercial da rede mundial de computadores,*]$_3$ [*pois ela torna possível realizar transações comerciais de forma global, com custo de manutenção inferior ao empregado em uma rede de comércio tradicional.*]$_4$

[*O objetivo deste trabalho é apresentar uma proposta para o projeto e implementação de um serviço de comércio eletrônico na plataforma JAMP.*]$_5$ [*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.*]$_6$

Figure 6. Text 2

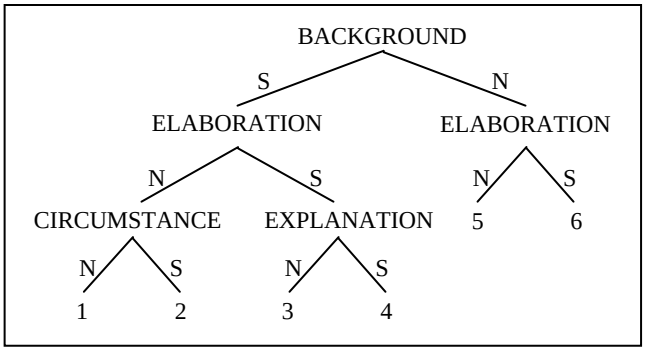

Figure 7. Text 2 rhetorical relation

The probability of the structure in Figure 7 is given by:

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 Evaluation

In order to objectively evaluate DiZer, a reference corpus was produced. The corpus, called Rhetalho (Pardo and Seno, 2005), is composed of 50 rhetorically annotated texts (with size around half a page) from scientific and news genres, which are not in

CorpusTCC. The scientific texts are from Computer Science domain; the news texts were collected from diverse sections from the on-line newspaper *Folha de São Paulo*.

All the texts were annotated by two judges (experts in RST), using Daniel Marcu's RST Annotation Tool (available at http://www.isi.edu/~marcu/discourse/) 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, modular and incremental; by modular, it means that clauses inside sentences must be related first, then sentences inside paragraphs must be related, and, finally, paragraphs must be related; by incremental, it means that, whenever possible, as soon as a new segment is determined, it must be related to the subtree already built until that point;
- only binary structures are allowed, i.e., each node in the tree may have 2 children at most; with this, when a non-binary tree is produced, it must be transformed in 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 by Carlson and Marcu (2001) must be followed (although they were defined for the English language, they are generic enough to be applied to Portuguese too); when the judges disagree about a segment, the most comprehensive 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.

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 nucleus 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 judge DiZer results validity, we run the same evaluation for a baseline method. The baseline method performs sentential segmentation and detects only ELABORATION relations (given that the ELABORATION relation is usually the most generic and frequent one in texts), with the first segment being the nucleus.

Table 1 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 2 presents the numbers for the news texts. F-measure (F), which is a combination of recall and precision, is also showed. It is a unique measure of how good a system is.

Table 1. DiZer performance for scientific texts

	DiZer – sentential segmentation (%)			DiZer – 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 2. DiZer performance for news texts

	DiZer – sentential segmentation (%)			DiZer – 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-measures, 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 DiZer bad results for sentential segmentation with news texts 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 works results, in particular, to Marcu's analyzer (1997, 2000), which is the most similar to DiZer in literature. Although this direct comparison is unfair, given that the languages and test corpora differ, it gives an idea of the state of the art results in cue-phrase-based analyzers.

In relation to the errors committed by DiZer, we identified some of the reasons that caused them. 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, discourse analysis may be compromised during clausal segmentation (if verbs are not correctly classified) and rhetorical relations

detection (when a discourse pattern 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 analyzed: in some cases, cue phrases are misused, which introduces errors during rhetorical relations detection.

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 happens frequently in automatic discourse analysis, given the subjectivity of texts) may avoid the construction of any structure. In addition to this, Pardo (2005) shows that it is possible to have plausible rhetorical structures even when the compositionality criterion is not applied.

Next section presents some conclusions and makes some final remarks.

5 Conclusion

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

Although DiZer was developed for scientific texts analysis, its evaluation shows that it is possible to achieve acceptable results for other text genres and domains. We believe that this happens because, in general, cue phrases are consistently used by people in any kind of text.

DiZer is the first step towards the automation of other levels of discourse analysis. As suggested by Pardo (2005), it is possible to map directly rhetorical relations to the semantic relations proposed by Kehler (2002). This should be investigated in the future.

Acknowledgments

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

References

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.

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

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

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

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

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.

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

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

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

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

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.

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.

Pardo, T.A.S. and Nunes, M.G.V. (2003). *A Construção de um Corpus de Textos Científicos em Português do Brasil e sua Marcação Retórica*. Technical Report N. 212. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, September, 26p.

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.

Pardo, T.A.S.; Nunes, M.G.V.; Rino, L.H.M. (2004). DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In the *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence* – SBIA (Lecture Notes in Artificial Intelligence 3171), pp. 224-234. São Luis-MA, Brazil. September, 29 - October, 1.

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.

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.

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

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

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*, Vol. 2, pp. 1133-1140. Tokyo, Japan.