DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese

Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes, Lucia Helena Machado Rino

Núcleo Interinstitucional de Lingüística Computacional (NILC)
CP 668 – ICMC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.nilc.icmc.usp.br
{thiago@nilc.icmc.usp.br; gracan@icmc.usp.br;
lucia@dc.ufscar.br}

Abstract. This paper presents DiZer, an automatic DIscourse analyZER for Brazilian Portuguese. Given a source text, the system automatically produces its corresponding rhetorical analysis, following Rhetorical Structure Theory – RST [1]. A rhetorical repository, which is DiZer main component, makes the automatic analysis possible. This repository, produced by means of a corpus analysis, includes discourse analysis patterns that focus on knowledge about discourse markers, indicative phrases and words usages. When applicable, potential rhetorical relations are indicated. A preliminary evaluation of the system is also presented.

Keywords: Automatic Discourse Analysis, Rhetorical Structure Theory

1 Introduction

Researches in Linguistics and Computational Linguistics have shown that a text is more than just a simple sequence of juxtaposed sentences. Indeed, it has a highly elaborated underlying discourse structure. In general, this structure represents how the information conveyed by the text propositional units (that is, the meaning of the text segments) correlate and make sense together.

There are several discourse theories that try to represent different aspects of discourse. The Rhetorical Structure Theory (RST) [1] is one of the most used theories nowadays. According to it, all propositional units in a text must be connected by rhetorical relations in some way for the text to be coherent. As an example of a rhetorical analysis of a text, consider Text 1 (adapted from [2]) in Figure 1 (with segments that express basic propositional units numbered) and its rhetorical structure in Figure 2. 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 CONTRAST relation, that is, they are opposing facts that may not happen at the same time; proposition 3 is the direct RESULT (non volitional) of the opposition between 1 and 2. In some cases, relations are multinuclear (e.g., CONTRAST relation), that is, they have no satellites and the connected propositions are considered to have the same importance.

[1] He wanted to play tennis with Jane, [2] but also wanted to have dinner with Susan. [3] This indecision drove him crazy.

Figure 1 – Text 1

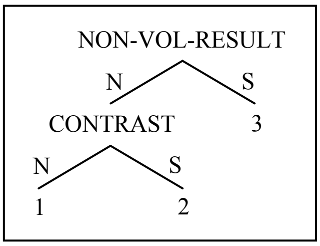

Figure 2 – Text 1 rhetorical structure

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 automatic text summarization (to identify the most important information of a text to produce its summary) (see, for instance, [2] and [3]), co-reference resolution (determining the context of reference in the discourse may help determining the referred term) (see, for instance, [4] and [5]), and for other natural language understanding applications as well. Some discourse analyzers are already available for both English and Japanese languages, (see, for example, [2], [6], [7], [8], [9], [22] and [23]).

This paper describes DiZer, an automatic DIscourse analyZER for Brazilian Portuguese. To our knowledge, it is the first proposal for this language. It follows those existing ones for English and Japanese, having as the main process a rhetorical analyzer, in accordance with RST. DiZer main resource is a rhetorical repository, which comprises knowledge about discourse markers, indicative phrases and words usages, and the rhetorical relations they may indicate, in the form of discourse analysis patterns. Such patterns were produced by means of a corpus analysis. When applied to an unseen text, they may identify the rhetorical relations between the propositional units. The rhetorical repository also comprises heuristics for helping determining some rhetorical relations, mainly those that are usually not superficially signaled in the text.

Next section presents some relevant aspects of other discourse analysis researches. Section 3 describes the corpus analysis and the repository of rhetorical information used in DiZer. Section 4 outlines DiZer architecture and describes its main processes. Section 5 shows some preliminary results concerning DiZer performance, while concluding remarks are given in Section 6.

2 Related Work

Automatic rhetorical analysis became a burning issue lately. Significant researches on such an issue have arisen that focus on different methodologies and techniques. This section sketches some of them.

Based on the assumption that cue-phrases and discourse makers are direct hints of a text underlying discourse structure, Marcu [6] was the first to develop a cue-phrase-based rhetorical analyzer for free domain texts in English. He used a corpus-driven methodology to identify discourse markers and information on their contextual occurrences and possible rhetorical relations. Marcu also proposed a complete formalization for RST in order to enable its computational manipulation according to his purposes. Later on, Marcu [2], Marcu and Echihabi [7] and Soricut and Marcu [8] proposed, respectively, a decision-based rhetorical analyzer, a Bayesian machine learning-based rhetorical analyzer and a sentence-level rhetorical analyzer using statistical models. In the first one, Marcu applied a shift-reduce parsing model to build rhetorical structures. He achieved better results than with the cue-phrase-based analyzer. In the second one, Marcu and Echihabi trained a Bayesian classifier only with the words of texts to identify four basic rhetorical relations. They achieved a high accuracy in their analysis. Finally, Soricut and Marcu made use of syntactic and lexical information extracted from discourse annotated lexicalized syntactic trees to train statistical models. With this method, in the sentence-level analysis, they achieved results near human performance.

Also based on Marcu's RST formalization, Corston-Oliver [9] developed a rhetorical analyzer for encyclopedic texts based on the occurrence of discourse markers in texts and syntactic realizations relating text segments. He investigated which syntactic features could help determining rhetorical relations, focusing on features like subordination and coordination, active and passive voices, the morphosyntactic categorization of words and the syntactic heads of constituents.

Following Marcu's analyzer [6], DiZer may also be classified as a cue-phrase-based rhetorical analyzer. However, differently from Marcu's analyzer, DiZer is genre specific. For this reason, it makes use of other knowledge sources (indicative phrases and words, heuristics) and adopts an incremental analysis method, as will be discussed latter in this paper. Next section describes the conducted corpus analysis for DiZer development.

3 Corpus Analysis and Knowledge Extraction

3.1 Annotating the Corpus

The corpus was composed of 100 scientific texts on Computer Science taken from the introduction sections of MsC. Dissertations (c.a. 53.000 words and 1.350 sentences). The scientific genre has been chosen for the following reasons: a) scientific texts are supposedly well written; b) they usually present more discourse makers and indicative phrases and words than other text genres; c) other works on discourse analysis for Brazilian Portuguese ([10], [11], [12], [13], [14]) have used the same sort of texts.

The corpus has been rhetorically annotated following Carlson and Marcu's discourse annotation manual [15]. Although this manual focuses on the English language, it may be also applied to Brazilian Portuguese, since RST rhetorical relations are 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 O'Donnel's RSTTool [16] was used. To guarantee consistency during the annotation process, the corpus has been annotated by only one expert in RST.

Initially, the original RST relations set has been used to annotate the corpus. When necessary, more relations have been added to the set. In the end, the full set amounts to 32 relations, as shown in Figure 3. The added ones are in bold face. Some of them (PARENTHETICAL and SAME-UNIT) are only used for organizing the discourse structure. The table also shows the frequency (in %) of each relation in the analyzed corpus.

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

Figure 3 – DiZer rhetorical relations set

The annotation strategy for each text was incremental, step by step, 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 of the text 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 very same reasoning is used in DiZer for analyzing texts. More details about the corpus and its annotation may be found in [17] and [18].

3.2 Knowledge Extraction

Once completely annotated, the corpus has been manually analyzed in order to identify discourse markers, indicative phrases and words, and heuristics that might indicate rhetorical relations. Based on this, discourse analysis patterns for each rhetorical relation have been yielded, currently amounting to 840 patterns. These convey the main information repository of the system.

As an example, consider the discourse analysis pattern for the OTHERWISE rhetorical relation in Figure 4. According to it, an OTHERWISE relation connects two propositional units 1 and 2, with 1 been 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,* (in English, 'or, alternatively,') be present in the beginning of the segment that expresses propositional unit 2.

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 discourse analysis patterns matches some portion of the text being processed, the corresponding rhetorical relation is supposed to occur between the appropriate segments.

The discourse analysis 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 indicative phrase composed by (1) a word whose lemma is *cujo* ('which', in English[1]), (2) followed by any word that indicates purpose (represented by the 'purWord' class, whose possible values are defined apart by the user), (3) followed by any adjective, (4) followed by a word whose lemma is *ser* (verb 'to be', in English). Based on similar features, any pattern may be represented. Complex patterns, possibly involving long distance dependencies, may also be represented by using a special character (*) to indicate jumps in the pattern matching process.

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

Figure 4 – Discourse analysis pattern for the OTHERWISE rhetorical relation

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

Figure 5 – Discourse analysis pattern for the PURPOSE rhetorical relation

For relations that are not explicitly signaled in the text, like EVALUATION and SOLUTIONHOOD, it has been possible to define some heuristics to enable the discourse analysis, given the specific text genre under focus. 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

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

4 DiZer Architecture

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 valid rhetorical structures. In what follows, each process is explained. Figure 6 presents the system architecture.

[1] Although 'which' is invariable in English, its counterpart in Portuguese, *cujo*, may vary in gender and number.

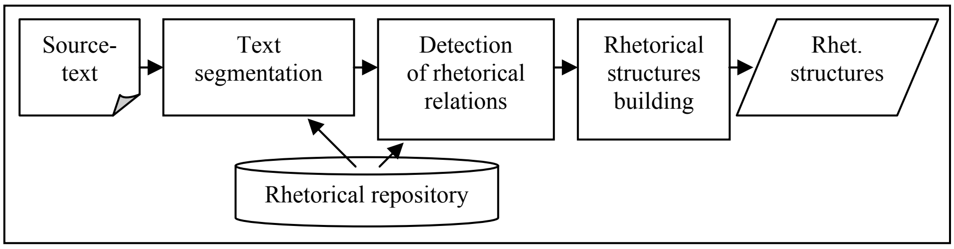

Figure 6 – DiZer architecture

4.1 Text Segmentation

In this process, DiZer tries to determine the simple clauses in the source text, since simple clauses usually express simple 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 [19]. Then, the segmentation process is carried out, segmenting the text always a punctuation signal (comma, dot, exclamation and interrogation points, etc.) or a strong discourse maker or indicative phrase is found. By strong discourse maker or indicative phrase we mean those words groups that unambiguously have a function in discourse. According to this, words like *e* and *se* (in English, 'and' and 'if'[2], respectively) are ignored, while words like *portanto* and *por exemplo* (in English, 'therefore' and 'for instance', respectively) are not. DiZer still verifies whether the identified segments are clauses by looking for occurrences of verbs in them.

Although this process is very simple, it produces reasonable results (see Figure 7 for an example of segmentation). In some cases, the system can not distinguish embedded clauses, causing inaccurate segmentation, but this may be overcome in the future by using a syntactic parser.

4.2 Detection of Rhetorical Relations

DiZer tries to determine at least one rhetorical relation for each two adjacent text segments representing the corresponding underlying propositions. In order to do so, it uses both discourse analysis patterns and heuristics. Initially, it looks for a relation between every two adjacent segments of each 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, as already discussed in this paper.

When more than one discourse analysis pattern apply, usually in occurrences of ambiguous discourse markers, all the possible patterns are considered. In this case, several rhetorical relations may be hypothesized for the same propositions. Because of this, multiple discourse structures may be derived for the same text.

[2] Although 'if' is rarely ambiguous in English, its counterpart in Portuguese, *se*, may assume many roles in a text. See a comprehensive discussion about *se* possible roles in [20].

In the worst case, when no rhetorical relation can be found between two segments, DiZer assumes a default heuristic: it adopts an ELABORATION relation between them, with the segment that appears first in the text being its nucleus. This is in accordance with what has been observed in the corpus analysis, in that the first segment is usually elaborated by following ones. Although this may cause some underspecification, or, maybe, inadequateness in the discourse structure, it is a plausible solution and it may even be the case that such relation really applies. ELABORATION was chosen as the default relation for being the most frequent relation in the corpus analyzed.

The system also keeps a record of the applied discourse analysis patterns and heuristics, so that it may be possible to identify later manually and/or computationally problematic/ambiguous cases in the discourse structure. In this way, it is possible to reengineer and improve the resulting discourse analysis.

4.3 Building the Rhetorical Structure

This step consists of determining the complete text rhetorical structure from the individual rhetorical relations between its segments. For this, the system makes use of the rule-based algorithm proposed in [6]. 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 [21]. When the final grammar is executed, all possible valid rhetorical structures are built.

As a complete example of DiZer processing, Figures 7 and 8 present, respectively, a text (translated from Portuguese) already segmented by DiZer and one of the valid rhetorical structures built. One may verify that the structure is totally plausible. It is also worth noticing that paragraphs and sentences form complete substructures in the overall structure, given the adopted processing strategy.

Next section presents some preliminary results concerning DiZer performance.

5 Preliminary Evaluation

A preliminary evaluation of DiZer has been carried out taking into account five scientific texts on Computer Science (which are not part of the corpus analyzed for producing the rhetorical repository). These have been randomly selected from introductions of MsC. dissertations of the NILC Corpus[3], currently the biggest corpora of texts for Brazilian Portuguese. Each text had, in average, 225 words, 7 sentences, 17 propositional units and 16 rhetorical relations.

Once discourse-analyzed by DiZer, the resulting rhetorical structures have been verified in order to assess two main points: (I) the performance of the segmentation process and (II) the plausibility of the hypothesized rhetorical relations. Such features have been chosen for being the core of DiZer main processes. Only one expert in RST has analyzed those structures, using as reference one manually generated discourse structure for each text, which incorporated all plausible relations between the propositions. Table 1 presents the resulting recall and precision average numbers for

[3] www.nilc.icmc.usp.br/nilc/tools/corpora.htm

DiZer. It also shows the results for a baseline method, which considers complete sentences as segments and always hypothesizes ELABORATION relations between them (since it is the most common and generic relation).

> [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 7 – Text 2

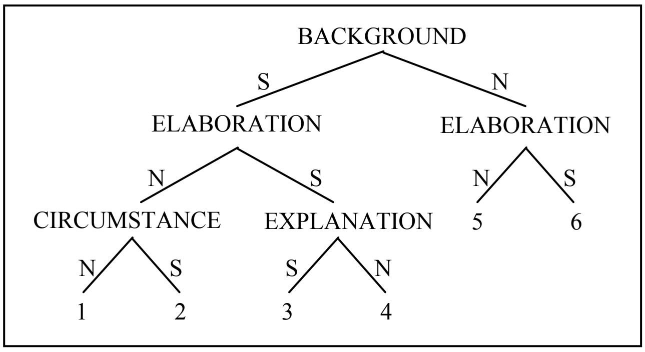

Figure 8 – Text 2 rhetorical structure

For text segmentation, recall indicates how many segments of the reference discourse structure were correctly identified and precision indicates how many of the identified segments were correct; for rhetorical relations hypotheses, recall indicates how many relations of the reference discourse structure were correctly hypothesized (taking into account the related segments and their nuclearity – which segments were nuclei and satellites) and precision indicates how many of the hypothesized relations were correct. It is possible to see that the baseline method performed very poorly and that DiZer outperformed it.

Table 1 – Evaluation results

	DiZer		**Baseline method**	
	Recall	*Precision*	*Recall*	*Precision*
Text segmentation	81%	87%	16%	32%
Rhet. relations hypotheses	71%	76%	22%	50%

Some problematic issues might interfere in the evaluation, namely, the tagger performance and the quality of the source texts. If the tagger fails in identifying the morphosyntactic classes of the words, discourse analysis may be compromised. Also, if the source texts present a significant misuse of discourse markers, inadequate rhetorical structures may be produced. These problems have not been observed in the current evaluation, but they should be taken into account in future evaluations.

It is worth noticing that Marcu's cue-phrase-based rhetorical analyzer (which is presently the most similar analyzer to DiZer), achieved worse recall in both cases (51% and 47%), but better precision (96% and 78%) than DiZer. Although this direct comparison is unfair, given that the languages, test corpora and even the analysis methods differ, it gives an idea of the state of the art results in cue-phrase-based automatic discourse analysis.

6 Concluding Remarks

This paper presented DiZer, a knowledge intensive discourse analyzer for Brazilian Portuguese that produces rhetorical structures of scientific texts based upon the Rhetorical Structure Theory. To our knowledge, DiZer is the first discourse analyzer for such language and, once available, must be the basis for the development and improvement of other NLP tasks, like automatic summarization and co-reference resolution.

Although DiZer was developed for scientific texts analysis, it is worth noticing that it may also be applied for free domain texts, since, in general, discourse markers are consistently used across domains.

In a preliminary evaluation, DiZer has achieved very good performance. However, there is still room for improvements. The use of a parser and the development of new specialized analysis patterns and heuristics must improve its performance. In the near future, a statistical module should be introduced into the system, enabling it to determine the most probable discourse structure among the possible structures built, as well as to hypothesize rhetorical relations in the case that there are not discourse markers and indicative phrases and words present in some segment in the source text.

Acknowledgments

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

References

1. Mann, W.C. and Thompson, S.A.: Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190 (1987).
2. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. The MIT Press. Cambridge, Massachusetts (2000).
3. O'Donnell, M.: Variable-Length On-Line Document Generation. In the Proceedings of the 6th European Workshop on Natural Language Generation. Duisburg, Germany (1997).
4. Cristea, D.; Ide, N.; Romary, L.: Veins Theory. An Approach to Global Cohesion and Coherence. In the Proceedings of Coling/ACL. Montreal (1998).
5. Schauer, H.: Referential Structure and Coherence Structure. In the Proceedings of TALN. Lausanne, Switzerland (2000).
6. Marcu, D.: The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Department of Computer Science, University of Toronto (1997).

7. Marcu, D. and Echihabi, A.: An Unsupervised Approach to Recognizing Discourse Relations. In the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA (2002).
8. Soricut, R. and Marcu, D.: Sentence Level Discourse Parsing using Syntactic and Lexical Information. In the Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Canada (2003).
9. Corston-Oliver, S.: Computing Representations of the Structure of Written Discourse. PhD Thesis, University of California, Santa Barbara, CA, USA (1998).
10. Feltrim, V.D.; Aluísio, S.M.; Nunes, M.G.V.: Analysis of the Rhetorical Structure of Computer Science Abstracts in Portuguese. In the Proceedings of Corpus Linguistics (2003).
11. Pardo, T.A.S. and Rino, L.H.M.: DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), Advances in Natural Language Processing, (2002) pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.
12. Aluísio, S.M. and Oliveira Jr., O.N.: A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users. Lecture Notes in Computer Science, Vol. 1010, (1995) pp. 121-132.
13. Aluísio, S.M.; Barcelos, I.; Sampaio, J.; Oliveira J, O.N.: How to Learn the Many Unwritten ´Rules of the Game´ of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases to Support Scientific Writing. In the Proceedings of the IEEE International Conference on Advanced Learning Technologies. Madison, Wisconsin. Los Alamitos, CA: IEEE Computer Society, Vol. 1, (2001) pp. 257-260.
14. Rino, L.H.M. and Scott, D.: A Discourse Model for Gist Preservation. In the Proceedings of the XIII Brazilian Symposium on Artificial Intelligence (SBIA'96). Curitiba - PR, Brasil (1996).
15. Carlson, L. and Marcu, D.: Discourse Tagging Reference Manual. ISI Technical Report ISI-TR-545 (2001).
16. O'Donnell, M.: RST-Tool: An RST Analysis Tool. In the Proceedings of the 6th European Workshop on Natural Language Generation. Gerhard-Mercator University, Duisburg, Germany (1997).
17. Pardo, T.A.S. e Nunes, M.G.V.: A Construção de um Corpus de Textos Científicos em Português do Brasil e sua Marcação Retórica. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, no. 212 (2003).
18. Pardo, T.A.S. e Nunes, M.G.V.: Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil. Relatório Técnico NILC-TR-04-03. Série de Relatórios do NILC, ICMC-USP (2004).
19. Aires, R.V.X.; Aluísio, S.M.; Kuhn, D.C.S.; Andreeta, M.L.B.; Oliveira Jr., O.N.: Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In the Proceedings of the Brazilian AI Symposium (SBIA'2000), (2000) pp. 20-22.
20. Martins, R.T.; Montilha, G.; Rino, L.H.M.; Nunes, M.G.V.: Dos Modelos de Resolução da Ambigüidade Categorial: O Problema do SE. In the Proceedings do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, PROPOR'99, (1999) pp. 115-128. Évora, Portugal.
21. Pereira, F.C.N. and Warren, D.H.D.: Definite Clause Grammars for Language Analysis – A Survey of the Formalism and Comparison with Augmented Transition Networks. Artificial Intelligence, N. 13, (1980) pp. 231-278.
22. Schilder, F.: Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), Natural Language Engineering, Vol. 8. Cambridge University Press (2002).
23. Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S.: A discourse structure analyzer for Japonese text. In the Proceedings of the International Conference on Fifth Generation Computer Systems, Vol. 2, (1992) pp. 1133-1140. Tokyo, Japan.