DiZer 2.0 – An Adaptable On-line Discourse Parser

Erick Galani Maziero[1], Thiago Alexandre Salgueiro Pardo[1], Iria da Cunha[2,3], Juan-Manuel Torres-Moreno[2,4], Eric SanJuan[2]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

[2] Traitement Automatique de la Langue Naturelle Ecrite (TALNE)
Laboratoire Informatique d'Avignon (LIA)
Université d'Avignon et des Pays de Vaucluse

[3] Grup Iulaterm
Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra

[4] Département de génie informatique
École Polytechnique de Montréal

[4] Grupo de Ingeniería Lingüística
Instituto de Ingeniería
Universidad Nacional Autónoma de México

egmaziero@gmail.com, taspardo@icmc.usp.br, iria.dacunha@upf.edu,
{juan-manuel.torres,eric.sanjuan}@univ-avignon.fr

Abstract. *This paper presents DiZer 2.0, an adaptable on-line discourse parser. It is an evolution of DiZer, the first version of the system for Brazilian Portuguese language. It keeps the same analysis method following the Rhetorical Structure Theory, but builds on it by allowing any user to run it on the web and, if necessary, to build its own parser by incorporating discourse knowledge of the desired language and text type/genre. Besides presenting the system main points, this paper also shows a case study, in which the system is adapted for parsing the Spanish language.*

1. Introduction

In the last decades, the Natural Language Processing (NLP) area has experienced incredible advances. From naïve and relatively simple resources and tools (e.g., the first thesauri and word-by-word machine translators) to very good applications (e.g., statistical machine translation – for which Google Translator is the most famous – and more intelligent information retrieval and extraction techniques – such as the ones used by WolframAlpha and Qwiki applications), it is not rare to attribute such success to the release and availability of large amounts of corpora and varied information/knowledge sources, as well as systems that process and produce them.

Anais do III Workshop "A RST e os Estudos do Texto", páginas 1–17, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

There are numerous examples of indispensable resources to NLP daily activities: Penn Treebank and its subsequent variations and improvements (Marcus et al. 1993; Marcus, 1994; Kingsbury and Palmer, 2002), Princeton Wordnet (Fellbaum, 1998) and its versions in several languages, Wikipedia, MIT Commonsense Computing Initiative (Liu, 2004; Silva et al., 2010), Framenet (Baker et al., 1998; Ruppenhofer et al., 2010), among several others. The same applies for tools, from simple to complex ones: stemmers (van Rijsbergen et al., 1980), syntactical parsers (Charniak, 1993; Collins, 1999; Atserias et al., 2006; Petrov and Klein, 2007), grammar checkers (Martins et al., 1998; Kinoshita et al., 2006), named entity taggers (Bikel et al., 1999; Cardoso, 2008), semantic parsers (Gildea and Jurafsky, 2002; Poon and Domingos, 2009), among several others. These varied linguistic levels of treatment have allowed the development of more intelligent and useful applications.

More recently, the discourse level has gained some prominence. Several works have explored traditional discourse models in computational applications as well as have showed that discourse parsers are also possible to exist with some minimum acceptable performance. The Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) is the discourse model by excellence used in the majority of NLP applications and tools, from text summarization, human and machine translation and opinion mining to essay scoring and coherence assignment (see, e.g., Marcu, 1997; Marcu et al., 2000; Pardo and Rino, 2002; Burstein et al., 2003; Somasundaran et al., 2009; da Cunha and Iruskieta, 2010; Uzêda et al., 2010). Parsers for some languages became available (see, e.g., Sumita et al., 1992; Marcu, 2000a; Pardo et al., 2008; Subba and Eugenio, 2009; among many others). Some of them, despite some limited assumptions, achieve near human performance (see, e.g., Soricut and Marcu, 2003, for sentence level analysis).

There are both symbolic and statistical approaches to discourse parsing. Some hybrid methods were also investigated. For Portuguese, the only discourse parser available is DiZer (Pardo and Nunes, 2008), a symbolic system that tries to map discourse cues in the text (as discourse markers and indicative words and phrases) into RST relations and, based on them, to build all valid discourse structures for the input text. DiZer is completely language dependent and customized for scientific texts.

Frequently it has been the case that a more generic discourse parser might be useful or that a parser for another language might also follow DiZer steps to perform the parsing. DiZer also showed to be a very heavy system (depending on several pre-existent tools) and difficult to adapt to different scenarios. Based on these facts, effort has been made to produce a new version of it, which is the focus of this paper.

This paper presents DiZer 2.0, an adaptable on-line discourse parser[1]. It keeps the same basic analysis method of its previous version (although simplifying some of them), but builds on it by allowing any user to run it on the web and, if necessary, to build its own parser by incorporating discourse knowledge of the desired language and text type/genre. Besides presenting the system main points, this paper also shows a case study, in which the system is adapted for parsing the Spanish language.

Next section briefly introduces RST. Section 3 describes the new system organization and its processes. Section 4 reports the case study of developing a Spanish discourse parser with DiZer 2.0. Finally, some final remarks are made in Section 5.

[1] Available at http://www.nilc.icmc.usp.br/dizer2/

2. Rhetorical Structure Theory

The RST was proposed by Mann and Thompson (1987) as a theory of text organization in terms of its propositions and their functions, i.e., how the adjacent propositions in the text – its discourse segments – relate to one another and provide the underlying intentions in the text regarding the writer (the producer of the text) purposes.

According to RST, propositions express basic meaningful units, usually expressed by clauses or sentences in a text. Their relationships are traditionally structured in a tree-like form (where larger units – composed by more than one proposition – are also related in the higher levels of the tree), although some recent works have argued that graphs are more suitable representations (see, e.g., Wolf and Gibson, 2005). Table 1 lists the original relations predicted by RST.

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

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

As an illustration, see the example below of two (numbered) clauses whose corresponding propositions are in a Concession relation (Mann and Thompson, 1987, p. 13):

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

New important relations were soon later included in this list by RST authors, as Means and List relations. Several other works have also created new relation sets, some shorter and other much longer than the original (see, e.g., Marcu, 1997), making the necessary adaptations for treating particular text genres and domains. Pardo (2005) defines for DiZer the relation set shown in Table 2, complementing the previous relation set with some relations from the work of Marcu (included in the fourth column of the table, e.g., explanation, attribution, parenthetical and same-unit).

Table 2 – Relation set defined by Pardo (2005)

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

It is interesting to see that, differently from the original RST relations, some relations defined by Marcu are of structural nature, i.e., they do not have a proper meaning, but are useful for connecting the constituent parts of propositions that for some reason were not textually adjacent. Parenthetical and Same-Unit relations are examples of this kind of relation.

RST also defines what is called nuclearity for each relation. The propositions in a relation are classified as nuclei or satellites: nuclei are more important propositions, while satellites are usually complementary information. Relations with one nucleus and one satellite are said to be mononuclear relations. Relations that only have nuclei – where all the propositions are equally important – are said to be multinuclear relations. Sequence, Contrast, List, Joint and Same-Unit are multinuclear relations; the others are mononuclear relations.

Figure 1 shows an example of a complete RST structure. In mononuclear relations, the arrows leave from the satellites and point to the nuclei, which are also indicated by a vertical line. One may see the hierarchical nature of this kind of structure, where there are relations among larger text units above the leaves level (e.g., the Elaboration relation, which connects the first segment with the following two segments).

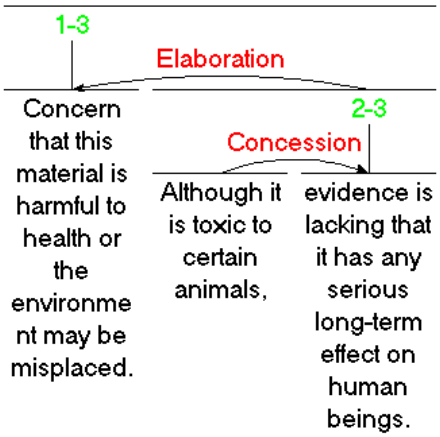

Figure 1 – Example of RST structure (Mann and Thompson, 1987, p. 15)

It is important to notice that RST analysis is highly dependent on the text understanding by the human that is performing the analysis. Therefore, it is usual that humans analyzing the same text may disagree in several aspects, from the definition of what constitutes the basic discourse segments to which relations hold among them and which segments are nuclei and satellites. In fact, it is acceptable that more than one RST structure may exist.

3. DiZer 2.0

The system starts by showing to the user two possible actions to perform: to start the discourse parsing or to manage the discourse knowledge repository, which is used to carry out the parsing. Figure 2 shows a screen dump of the system when it is loaded for the first time.

We will start by the parsing process itself. Following its first version, parsing in DiZer 2.0 is composed by three main steps: text segmentation, detection of rhetorical relations, and building of rhetorical structures.

Text segmentation may be manually or automatically performed, according to the user desire. If the user decides for the manual segmentation, the system will offer him a text box where the text must be inserted and segmented. To indicate the segments, the user must put each segment in a new line and, when sentences and paragraphs boundaries are found, they must be indicated by the [s] and [p] marks, respectively.

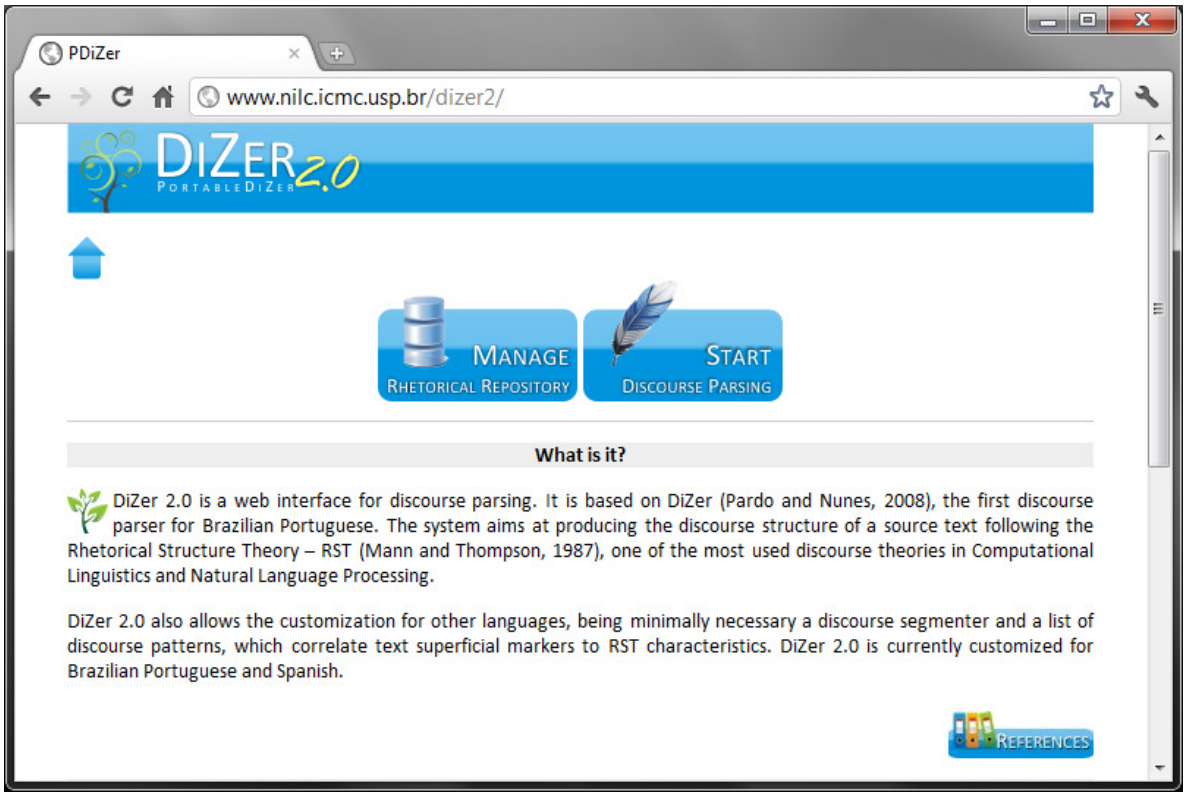

Figure 2 – DiZer 2.0 initial screen dump

Suppose that the following text (with 1 paragraph and 2 sentences) is under analysis (it is written in Brazilian Portuguese, but its English version is also shown):

> *Os resultados da análise de inteligibilidade de córpus podem ajudar a guiar a tarefa de simplificação textual, fornecendo quais características realmente tornam um texto mais simples de ser entendido por pessoas dos mais diversos níveis de letramento. Apesar de ter sido criada para este fim, a ferramenta pode ser utilizada para quaisquer fins que necessitem de tais informações.*
>
> ENGLISH TRANSLATION. The results of the corpus readability analysis may help guiding the text simplification task, providing the characteristics that really make a text easier to understand by people of varied literacy levels. Despite having been created for this purpose, the tool may be used for any purposes that require such information.

The manually segmented text would be the following:

> *Os resultados da análise de inteligibilidade de córpus podem ajudar a guiar a tarefa de simplificação textual,*
>
> *fornecendo quais características realmente tornam um texto mais simples de ser entendido por pessoas dos mais diversos níveis de letramento.* [s]
>
> *Apesar de ter sido criada para este fim,*
>
> *a ferramenta pode ser utilizada para quaisquer fins que necessitem de tais informações.* [s][p]
>
> ENGLISH TRANSLATION. The results of the corpus readability analysis may help guiding the text simplification task,
>
> providing the characteristics that really make a text easier to understand by people of varied literacy levels. [s]
>
> Despite having been created for this purpose,
>
> the tool may be used for any purposes that require such information. [s][p]

One may see that the segmentation results in 4 segments. Although in this example it was adopted the clause segmentation, the user may use the segmentation he judges the most appropriate, e.g., sentence or paragraph segmentation. Figure 3 shows a screen dump of the manual segmentation interface.

Sentences and paragraphs boundaries must be marked in order to allow DiZer 2.0 to perform what has been called incremental analysis (during the next step – detection of rhetorical relations). In this analysis style, the system tries to take advantage of the text organization produced by its writer. It assumes that adjacent clauses inside sentences must be related first. Then, it tries to relate adjacent sentences inside paragraphs. Finally, adjacent paragraphs are related. Since language use is almost unrestricted and several writing styles exist, such incremental analysis may not always apply, but it is undoubtedly an interesting analysis criterion, being also useful for restricting the number of possible analyses.

If the user decides for the automatic segmentation, then the appropriate segmentation tool must be made accessible to DiZer 2.0, either by being directly included in it (with the support of the system development team) or by being remotely called by the system, assuming the form of a web service. For Portuguese language, DiZer 2.0 directly incorporates a segmentation tool that makes use of a syntactical parser – PALAVRAS parser (Bick, 2000) – in order to identify clauses, which are the segments usually considered for RST analysis. On the other hand, the DiZer 2.0 adaptation to the Spanish language makes use of a segmentation tool called through web. This tool, in turn, also uses a parser for Spanish. Such adaptation is introduced in the next section. It is important to notice that any resource or tool that may be necessary to the segmentation may be used in DiZer 2.0 environment.

The used segmentation tool must produce as output the same text format of the manual segmentation, i.e., one segment per line and the sentence and paragraph boundaries marks. Otherwise, DiZer 2.0 wil have to adapt this output to the appropriate format. This is how it works for Spanish today, since the Spanish segmentation tool produces a different XML output format.

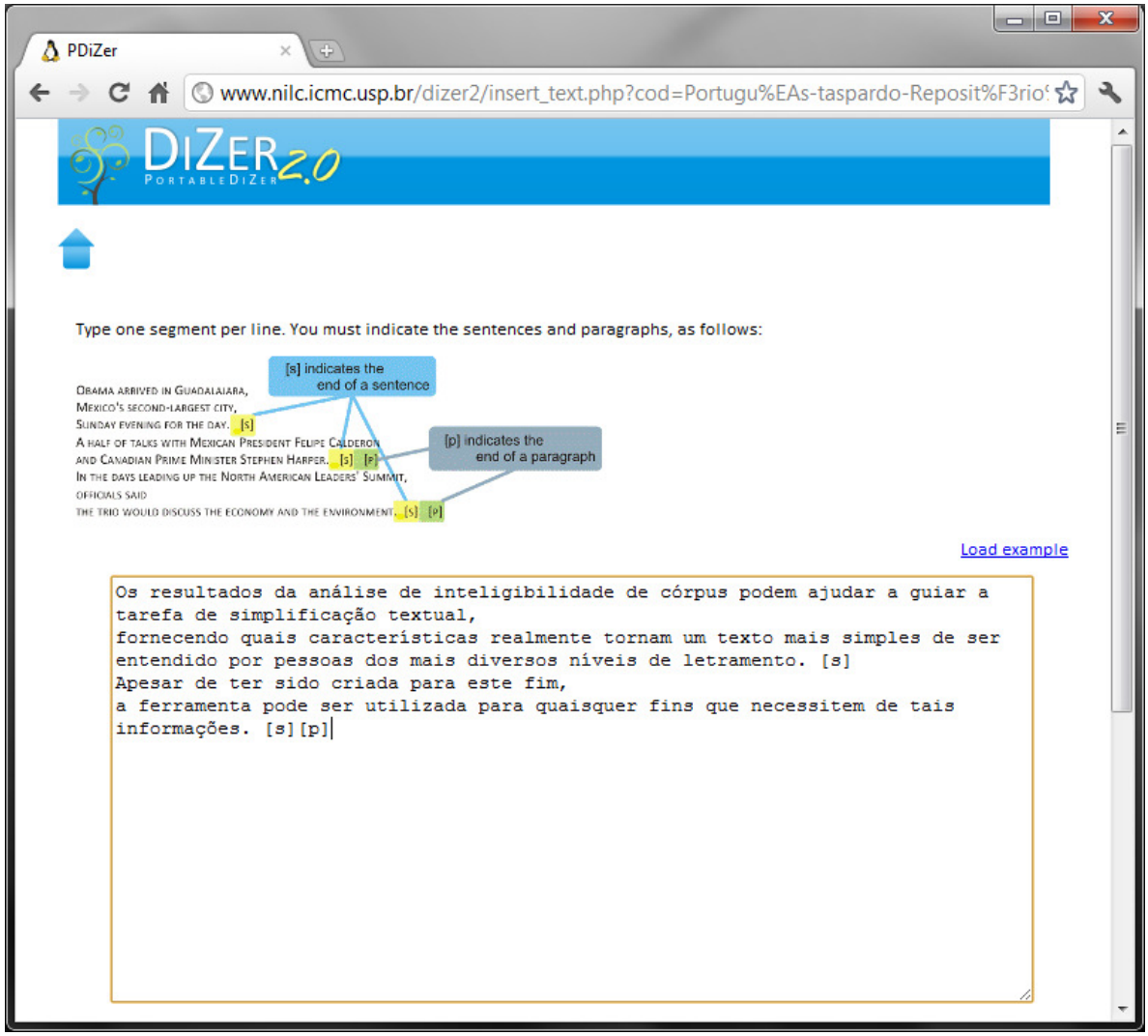

Figure 3 – Dump of the manual segmentation screen

After segmentation is done, DiZer 2.0 proceeds to detect the rhetorical relations among pairs of segments, respecting the incremental analysis style cited above. This step is purely symbolic and consists of predicting possible relations given some text hints/cues/marks, as discourse markers and indicative words and phrases that happen in the pair of segments. For instance, it is well known that the discourse markers "but" and "therefore" indicate opposition (e.g., contrast, antithesis or concession) and cause-effect (e.g., vol. cause, non-vol. cause, vol. result or non-vol. result) relations, respectively. The same happens for special words and phrases, such as "performance" and "the purpose of this work is…", which might indicate, for instance, evaluation and purpose relations. It is also well known that these correspondences are not deterministic, since discourse markers and indicative words and phrases may indicate more than one relation and that each relation may be signaled by different discourse markers. Relations may also happen without any marks.

DiZer 2.0 stores the correspondences among the RST relations and the text marks in the form of discourse patterns/templates. Figure 4 shows a discourse pattern for the concession relation.

Relation/pattern	Concession
Order	SN
Marker 1	---
Position of Marker 1	---
Marker 2	but
Position of Marker 2	Beginning

Figure 4 – A discourse pattern for detecting the concession relation

This pattern says that a concession relation will be detected between two segments when there is the word "but" in the beginning of the second segment, and that the first segment will be the satellite of the relation, while the second will be the nucleus (indicated by the SN information in the "order" field). Notice that the information about the marker in the first segmented is left not specified, given that it is not used in this case.

Figure 5 shows a more complex pattern for detecting a sequence relation. Notice that the relation will be detected if the first segment has the word "first" in any position and the second segment has the word "then" in its beginning. As the sequence relation is multinuclear, both segments will be classified as nuclei (indicated by NN).

Relation/pattern	Sequence
Order	NN
Marker 1	first
Position of Marker 1	Any
Marker 2	then
Position of Marker 2	Beginning

Figure 5 – A discourse pattern for detecting the sequence relation

Any configuration of discourse pattern is possible. Other information may also be included in the patterns, e.g., part-of-speech tags and lemmas, in order to compose more complex markers. For instance, Figure 6 says that, for detecting a purpose relation, it is necessary to find in the second segment an indicative phrase composed by the word "the", a word of purpose type, the word "of", the word whose lemma is "this", a word whose part-of-speech tag is "noun", and a word whose tag is "verb".

Relation/pattern	Purpose
Order	NS
Marker 1	---
Position of Marker 1	---
Marker 2	the purposeWord_list of this_lem _noun _verb
Position of Marker 2	Beginning

Figure 6 – A discourse pattern for detecting the purpose relation

The list of words that are of purpose type must be specified separately and, in the pattern, must be indicated by the _list mark. Part-of-speech tags must come in the _tag format, while lemma information must be indicated by using the _lem mark. The list of purpose words might contain, for instance, the words "purpose" and "aim".

The discourse patterns also accept the specification of optional words (by using the _opt mark after a word) and non-contiguous marks, which is indicated by the '*' symbol. For instance, Figure 7 shows a hypothetical pattern for the evaluation relation. It says that an evaluation relation occurs when the second segment has the word "performance" followed (not necessarily adjacent) by the optional word "very" and the word "good".

Relation/pattern	Evaluation
Order	NS
Marker 1	---
Position of Marker 1	---
Marker 2	performance * very_opt good
Position of Marker 2	Any

Figure 7 – A discourse pattern for detecting the evaluation relation

DiZer 2.0 offers the user an interface for defining all the necessary discourse patterns, as well as the word lists that are necessary to build the patterns. This interface is found by following DiZer 2.0 initial option to manage the discourse knowledge repository. The interface for defining discourse patterns is illustrated in Figure 8. Figure 9 shows the interface for defining a word list.

Figure 8 – Dump of the discourse pattern definition screen

In general, each relation will have a large set of corresponding patterns. The first version of DiZer (Pardo and Nunes, 2008) counted with more than 750 patterns, which were designed based on a corpus study. DiZer 2.0 also offers facilities for visualizing the patterns and lists, to alter pre-existent patterns, and to import previously defined patterns.

After performing the segmentation, the user must specify in DiZer 2.0 interface the language and the repository of discourse patterns that he desires to use. Then, using the selected repository, DiZer 2.0 performs the detection of rhetorical relations by simply matching all the patterns with the pairs of segments. All the relations that are identified are stored. To do the matching, DiZer 2.0 generates a set of regular expressions from the discourse patterns, and then applies such expressions to the segments.

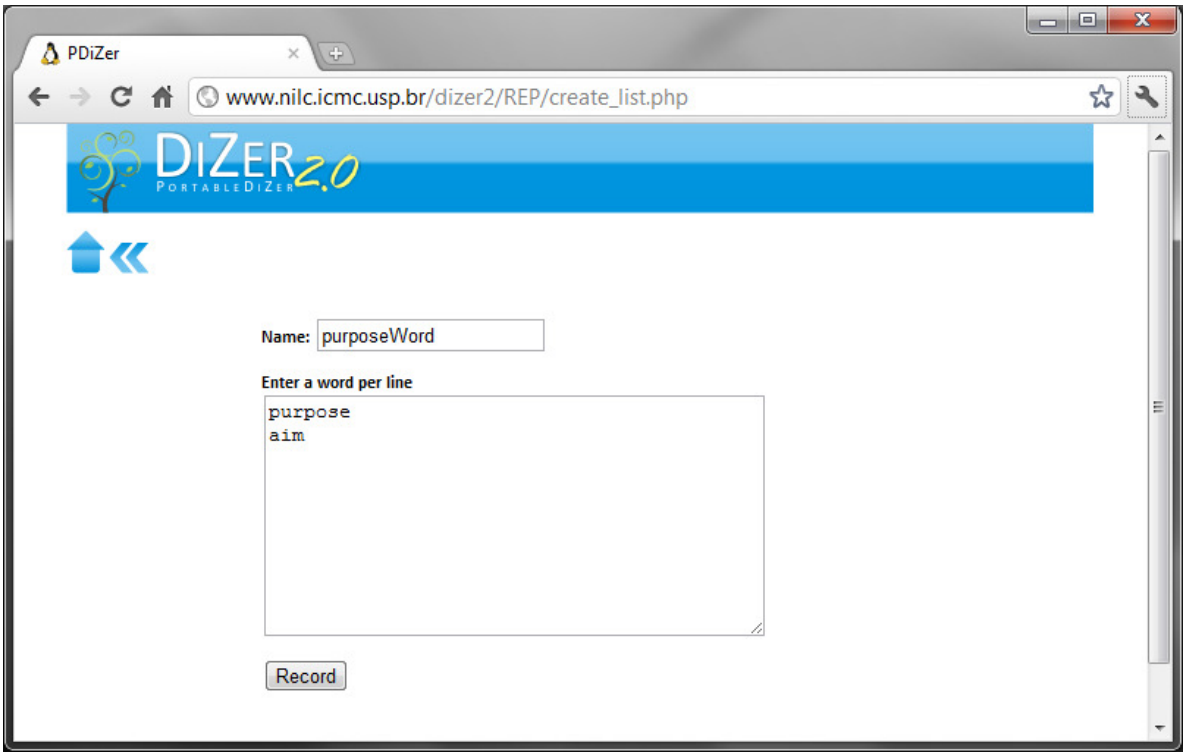

Figure 9 – Dump of the word list definition screen

Since discourse patterns may use part-of-speech and lemma information, such information must be previously provided. It may come from a tagger and a lemmatizer, or from a syntactical parser. This step of enriching the input text with this information may be done during segmentation or during the detection of relations, before doing the matching. In fact, each word in the segments under analysis must be replaced by a triple in the following format: word(lemma)_tag. When lemmas and tags are not necessary, such information may be left not specified. The use of such information depends solely on the design of the discourse patterns. For Portuguese and Spanish languages, syntactical parsers are used to provide such information. If not already used in the segmentation step, the parser may be inserted into DiZer 2.0 (with the support of the system development team).

The last step in DiZer 2.0 consists in producing the final RST trees from the relations detected before for each segment pair. For this end, it simply follows the strategy proposed by Marcu (1997).

The construction of the final RST trees may be influenced by some parameters that the user specifies. The user may decide to consider the nuclei restriction, according to which a relation that connects two subtrees must connect mainly their nuclei. Although this restriction predicts the construction of valid and well-formed trees, it may

be too strong for automatic purposes, since it may be difficult to observe it for every case, what might prevent that any trees are built for some texts.

Other option that the user has is to join trees of similar structure. For instance, suppose that the root of a tree might have an elaboration or an evidence relation. If the user decides not to use the join option, two RST trees will be produced, each one with one of the two possible relations; if the user decides to use the join option, only one RST tree will be produced, and its root will be labeled with the two relations, indicating that any of them might apply in this case. This option is useful when the user does not want to produce a large set of trees, but, instead, would like to see the structural variety that might exist for a text.

After DiZer 2.0 performs the complete RST parsing, it shows the final RST trees in logical and tree-like format. It also exhibits all the intermediary data produced during the parsing, as the segments that were found, the discourse patterns that were applied and the identified relations, and the runtimes, among other information. Figure 10 illustrates the output of DiZer 2.0 process for the example text in the beginning of this section. One may see that the system produced 2 possible trees, and that the first one looks to be the most appropriate analysis for the text.

Figure 10 – Dump of the output screen

To use the system fully (including creating and managing a repository of discourse patterns), the user needs to log in the system. Once logged, all of its information is

permanently stored (unless the user decides to erase it). Such mechanism also incorporates some security for the work the user is carrying out.

4. The Development of a Spanish Discourse Parser

Nowadays, the adaptation of DiZer 2.0 to Spanish is being carried out. A Spanish discourse segmenter has been developed (da Cunha et al., 2010a, 2010b), which has been integrated in the DiZer 2.0 interface. This segmenter, called DiSeg, has been evaluated using as gold standard a corpus including medical texts (obtaining an 80% of F-score) and terminological texts (obtaining a 91% of F-score). This segmenter and the gold standard are on-line[2]. DiSeg is based on a set of discourse segmentation rules using lexical and syntactic features. These rules are based on: discourse markers, as "because"' (porque), "nevertheless" (sin embargo) or "in summary" (en resumen), which usually mark relations of Cause, Antithesis and Summary, respectively; conjunctions, for example, "or" (o) or "but" (pero); adverbs, as "anyway" (de todas maneras); verbal forms, as gerunds, finite verbs, etc.; and punctuation marks, as parenthesis or dashes. DiSeg implementation relies on the shallow parsing of Freeling (Atserias et al., 2006). One example of rule is the following one:

> IF one or more verbs (finite, infinitive or gerund) are detected into a sentence AND afterwards a discourse marker is detected AND after this discourse marker there is another verb or verbs (finite, infinitive or gerund) THEN insert a segment boundary before the discourse marker

Following this rule, the next example would be segmented as it is indicated with brackets:

> [*También hay que recordar que el estudio se ha llevado a cabo en mujeres visitadas en el Hospital Clínic de Barcelona,*] [*por lo tanto la extrapolación de los resultados a la población general, incluso a aquellas mujeres residentes en la ciudad de Barcelona, puede producir sesgos.*]
>
> ENGLISH TRANSLATION. [It also has to be remember that the study has been carried out over women visited in the Clinic Hospital of Barcelona,] [therefore the extrapolation of these results to the general population, even to those women resident in Barcelona city, can produce biases.]

In order to build the rhetorical repository, the RST Spanish Treebank has been developed (da Cunha et al., 2011). This is the first corpus for Spanish annotated with RST relations. It includes 52,746 words, 267 texts, 2,256 sentences and 3,349 EDUs. It contains specialized texts from 9 domains: Astrophysics, Earthquake Engineering, Economy, Law, Linguistics, Mathematics, Medicine, Psychology and Sexuality. It includes 69% of texts annotated by one person (reference corpus for the rhetorical repository building) and 31% of double annotated texts (test corpus for the parser evaluation), following the methodology of the RST Discourse Treebank (Carlson et al., 2002a, 2002b). The rhetorical relations annotated are the same ones that those used for the development of the Spanish parser (see Table 3).

[2] Available at http://daniel.iut.univ-metz.fr/DiSeg/

Table 3 – Relation set used for the Spanish parser

Circumstance	Justification	Condition	Contrast
Solutionhood	Cause	Otherwise	Joint
Elaboration	Motivation	Interpretation	Means
Background	Result	Evaluation	List
Enablement	Purpose	Reformulation	Conjunction
Motivation	Antithesis	Summary	Unless
Evidence	Concession	Sequence	Same-Unit

Some examples of the information in the current discourse patterns under development are included in Table 4.

Table 4 – Some examples of linguistic patterns for relation detection

Relations	**Discourse patterns**
Cause	NUCLEUS + "because" (porque) + SATELLITE Ex. [Compré pollo] [porque tenía hambre.] [I bought chicken] [because I was hungry.]
	NUCLEUS + "since" (ya que) + SATELLITE Ex. [Compré pollo] [ya que tenía habmbre.] [I bought chicken] [since I was hungry.]
	"since" (ya que) + SATELLITE + NUCLEUS Ex. [Ya que tenía hambre,] [compré pollo.] [Since I was hungry,] [I bought chicken.]
Purpose	NUCLEUS + "in order to" (para) + SATELLITE Ex. [Tomé un taxi] [para llegar rápido.] [I took a taxi] [in order to arrive quickly.]
	"in order to" (para) + SATELLITE + NUCLEUS Ex. [Para llegar rápido,] [tomé un taxi.] [In order to arrive quickly,] [I took a taxi.]
Antithesis	NUCLEUS + "however" (sin embargo) + SATELLITE Ex. [Me gusta la carne.] [Sin embargo no me gusta el cerdo.] [I love meat.] [However I don't like pork.]
	NUCLEUS + "nevertheless" (no obstante) + SATELLITE Ex. [Me gusta la carne.] [No obstante no me gusta el cerdo.] [I love meat.] [Nevertheless I don't like pork.]
Reformulation	NUCLEUS + "in other words" (en otras palabras) + SATELLITE Ex. [Me gustan las manzanas, las naranjas, el melón, etc.] [En otras palabras, me encanta la fruta.] [I like apples, oranges, melon, etc.] [In other words, I love fruit.]
	NUCLEUS + "that is" (es decir) + SATELLITE Ex. [Me gustan las manzanas, las naranjas, el melón, etc.,] [es decir, me encanta la fruta.] [I like apples, oranges, melon, etc.,] [that is, I love fruit.]

At the moment, in order to build an exhaustive list of patterns for Spanish, the corpus analysis is being carried out. To do it, the RST Toolkit[3] is being used. This tool is

[3] Available at http://www.icmc.usp.br/~taspardo/Projects.htm

designed to include RST corpora previously annotated with the annotation tool RSTtool[4] (O'Donnell, 2000) or the ISI RST Annotation Tool[5], an extension of RSTtool. The Rhetorical Database (a module of the toolkit) allows the user to annotate the lexical markers relating discourse elements (nuclei or satellites) with their corresponding positions (before, after or in the middle of the segments).

Once the final rhetorical repository is compiled, it will be included in the DiZer 2.0 interface, using the interface designed for this task. At the moment, a beta version of the rhetorical repository is done, including 51 lexical markers corresponding to 19 relations. Nowadays, the DiZer 2.0 system allows the building of rhetorical trees in Spanish automatically. Nevertheless, as we have said, the actual version of the system is a beta version, so the tree building has yet limitations (not all the RST relations are shown and not all the connectors are detected).

Our short-term aim is to finish the rhetorical repository for Spanish and to evaluate the Spanish discourse parser (we call it ADAe). To do it, we will use the double annotated corpus of the RST Spanish Treebank. To evaluate the results of ADAe, we will use the RSTeval tool (Mazeiro and Pardo, 2009). This is an on-line tool for the automatic comparison of RST rhetorical trees (human and automatic built trees), following the methodology of Marcu (2000b). This methodology evaluates the similarity of simple segments, spans of more than one segment (in the higher levels of the tree), nuclearity and relations between two rhetorical trees, using traditional precision and recall measures.

5. Final Remarks

DiZer 2.0 is constantly under improvement. Currently, we are looking for faster ways of building the final RST trees and for selecting the trees to show to the final user. Better tree visualization methods are also under investigation. The system has some limitations by being purely symbolic, but altering it to deal with statistics and machine learning in an easy to use fashion may also be a future research challenge.

We hope that DiZer 2.0 is a useful platform for developing new discourse parsers or adapting existent ones for new text genres/domains.

Acknowledgments

The authors are grateful to FAPESP, CAPES and CNPq for supporting this work. This research is partially supported by: a postdoctoral grant (National Program for Mobility of Research Human Resources; National Plan of Scientific Research, Development and Innovation 2008-2011) given to Iria da Cunha by Ministerio de Ciencia e Innovación (Spain), and the research projects RICOTERM (FFI2010-21365-C03-01) and APLE (FFI2009-12188-C05-01).

References

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L.; Padró, M. (2006). FreeLing 1.3. Syntactic and semantic services in an open-source NLP library. In N. Calzolari et al. (Eds.), *Proceedings of the 5th International Conference on Language*

[4] Available at http://www.wagsoft.com/RSTTool/
[5] Available at http://www.isi.edu/~marcu/discourse/

Resources and Evaluation (LREC'06), pp. 48-55. Genoa, Italy: European Language Resources Association (ELRA).

Baker, C.F.; Fillmore, C.J.; Lowe, J.B. (1998). The Berkeley FrameNet project. In the *Proceedings of the COLING-ACL*.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis. Aarhus University. Denmark University Press.

Bikel, D.M.; Schwartz, R.; Weischedel, R.M. (1999). An Algorithm that Learns What's in a Name. *Machine Learning*, Vol. 34, pp. 211-231.

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

Cardoso, N. (2008). REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In C. Mota and D. Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pp. 195-211.

Carlson, L.; Marcu, D.; Okurowski, M.E. (2002a). *RST Discourse Treebank*. Pennsylvania: Linguistic Data Consortium.

Carlson, L.; Marcu, D.; Okurowski, M.E. (2002b). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In the *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge: MIT Press.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD Dissertation, University of Pennsylvania.

da Cunha, I. and Iruskieta, M. (2010). Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies*, Vol. 12, N. 5, pp. 563-598.

da Cunha, I.; SanJuan, E.; Torres-Moreno, J.M.; Lloberes, M.; Castellón, I. (2010a). Discourse Segmentation for Spanish based on Shallow Parsing. *Lecture Notes in Computer Science*, Vol. 6437, pp. 13-23.

da Cunha, I.; SanJuan, E.; Torres-Moreno, J.M.; Lloberes, M.; Castellón, I. (2010b). DiSeg: Un segmentador discursivo automático para el español. *Procesamiento del Lenguaje Natural*, Vol. 45, pp. 145-152.

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

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, Vol. 28, N. 3, pp. 245-288.

Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In the *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Kinoshita, J.; Salvador, L.N.; Menezes, C.D. (2006). CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus. In the *Proceedings of LREC*.

Liu, H. (2004). Commonsense reasoning in and over natural language. In the *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems*.

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

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

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

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

Marcu, D.; Carlson, L.; Watanabe, M. (2000). The Automatic Translation of Discourse Structures. In the *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 9-17.

Marcus, M.; Santorini, B.; Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, Vol. 19, N. 2, pp. 313-330.

Marcus, M. (1994). The Penn Treebank: A revised corpus design for extracting predicateargument structure. In the *Proceedings of the ARPA Human Language Technology Workshop*.

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*, Vol. 4, pp. 287-307. Cambridge University Press.

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

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

Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *3rd International Conference: Portugal for Natural Language Processing* – PorTAL (Lecture Notes in Artificial Intelligence 2389), pp. 263-273.

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

Pardo, T.A.S. and Nunes, M.G.V. (2008). On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43-64.

Petrov, S. and Klein, D. (2007). Improved Inference for Unlexicalized Parsing. In the *Proceedings of HLT-NAACL.*

Poon, H. and Domingos, P. (2009). Unsupervised semantic parsing. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1-10.

Ruppenhofer, J.; Ellsworth, M.; Petruck, M.R.L.; Johnson, C.R.; Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*.

Silva, M.A.R.; Dias, A.L.; Anacleto, J.C. (2010). Processing Common Sense Knowledge to Develop Contextualized Computer Applications. In the Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems.

Somasundaran, S.; Namata, G.; Wiebe, J.; Getoor. L. (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity

classification. In the *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 170-179.

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

Subba, R. and Di Eugenio, B. (2009). An effective discourse parser that uses rich linguistic information. In the *Proceedings of HLT-ACL*, pp. 566-574.

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.

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

van Rijsbergen, C.J.; Robertson, S.E.; Porter, M.F. (1980). *New models in probabilistic information retrieval*. British Library Research and Development Report, no. 5587. London: British Library.

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

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

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

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

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

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

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

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

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

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