A Symbolic Corpus-based Approach to Detect and Solve the Ambiguity of Discourse Markers

Iria da Cunha

University Institute for Applied Linguistics
Universitat Pompeu Fabra
C/ Roc Boronat, 138, 08018, Barcelona, Spain
iria.dacunha@upf.edu

Abstract. At present, discourse parsing is an important research topic. Rhetorical Structure Theory (RST) is one of the most popular approaches in this field. In general, discourse parsing includes three stages: discourse segmentation, discourse relations detection and building up rhetorical trees. Different strategies are used when developing discourse parsers. One of the strategies to detect discourse relations is based on symbolic rules that take into account linguistic clues, such as discourse markers. Nevertheless, some discourse markers are ambiguous, that is, they can indicate more than one discourse relation. This fact constitutes a problem when assigning discourse relations automatically. In this paper, a symbolic approach to detect and solve discourse markers ambiguity in Spanish is developed. First, we detect ambiguous discourse markers, using the training corpus of the RST Spanish Treebank. Second, we extract linguistic contexts for these markers. Third, we design linguistic rules to solve the ambiguity of discourse markers. Fourth, we evaluate the rules, using the test corpus of the RST Spanish Treebank. Our approach outperforms the baseline created following the methodology of the state of the art. Therefore, we consider that the results obtained in our experiments are representative and constitute the first step towards the disambiguation of discourse markers senses in Spanish. However, there is room for improvement and the main limitations of the approach are presented. In the future, the rules will be integrated in a discourse parser for Spanish, and several related applications will be developed (automatic summarization and information extraction, among others).

Keywords: Discourse Parsing, Discourse Markers, Ambiguity, Corpus, Rhetorical Structure Theory, Spanish

1 Introduction[1]

At present, discourse parsing is an important research topic, since it is being widely used to develop several applications, such as automatic summarization, information extraction, text generation, automatic translation, sentence compression, coherence evaluation, etc. Rhetorical Structure Theory (RST) [1] is one of the most popular

[1] This work has been financed by the Spanish projects RICOTERM 4 (FFI2010-21365-C03-01) and APLE 2 (FFI2012-37260), and a *Juan de la Cierva* grant (JCI-2011-09665).

approaches in this field. RST is a language-independent theory based on the idea that a text can be segmented into Elementary Discourse Units (EDUs) linked by means of nucleus-satellite or multinuclear rhetorical relations. In the first case, the satellite gives additional information about the other unit (the nucleus), on which it depends (e.g. Cause, Purpose or Result). In the second case, several elements, all nuclei, are connected at the same level, i.e. there are no dependent elements and they all are equally important with regard to the author's intentions (e.g. List, Contrast or Sequence). In general, discourse parsing includes three stages: discourse segmentation, discourse relations detection and building up rhetorical trees. RST-based discourse parsers for some languages are available: English [2], [3], [4], Japanese [5], Brazilian Portuguese [6] and Spanish [7]. These parsers use symbolic or statistical approaches. One of the strategies to detect discourse relations is based on symbolic rules which take into account linguistic clues, such as discourse markers [6], [8]. Traditionally, discourse markers are defined as invariable linguistic units that guide inferences in communication (see [9] for a review on discourse markers definitions). However, as [9] mentions: "[…] the signalling of discourse relations is not restricted to discourse markers; many other devices are used to signal the presence of such relations". Thus, we do not follow the traditional definition of discourse markers, but we use this term in a wide sense.

It is important to highlight that some discourse markers are ambiguous. Specifically, as [10] state:

> There are two types of ambiguity that need to be resolved during discourse processing. First, a word can be ambiguous between discourse or non-discourse usage. For example, 'once' can be either a temporal discourse connective or a simply a word meaning 'formerly'. Secondly, some connectives are ambiguous in terms of the relation they mark. For example "since" can serve as either a temporal or causal connective.

In this work, we focus on the second type of ambiguity. As [11] states, one of the problems of the semantics of natural connectors is that the same connector can express different connection types and one connection type can be expressed by several connectors. For example, the Spanish marker *mientras* ("while") can express at the same time Contrast, Circumstance and Condition. Regarding discourse parsing, [2] highlights that discourse markers can indicate more than one discourse relation and this fact constitutes a problem when assigning discourse relations automatically. When working on discourse parsing (specifically, in the case of automatic relation detection), three strategies can be used to deal with the problem of markers ambiguity: a) to choose the relation with a higher number of markers of this type, b) to give to the algorithm all possible relations, or c) to develop more fine-grained strategies combining several markers to choose only one relation.

The main objectives of this paper are: a) to detect ambiguous discourse markers in Spanish, and b) to develop fine-grained strategies in order to solve the ambiguity of discourse markers automatically. This is the first study that aims at detecting and solving the ambiguity of discourse markers in Spanish, considering ambiguity as the possibility to indicate more than one discourse relation.

In Section 2, related work is presented. In Section 3, the methodology used in the study is explained. In Section 4, corpus analysis and results are presented. In Section 5, the evaluation of the results is shown. In Section 6, some conclusions and future work are established.

2 Related Work

Most of the work on discourse markers disambiguation has been done for English. [12] carry out an empirical study on discourse and sentential uses of cue phrases, in which text-based and prosodic features are examined for disambiguation power. They propose that discourse or sentential usage can be distinguished by intonational features, and present a prosodic model that characterizes these distinctions. [13] present a set of manual sense annotation studies for three connectives in English ("since", "while" and "then"), whose arguments have been annotated in the Penn Discourse Treebank (PDTB) [14]. They use syntactic features annotated in this corpus and a maximum entropy model to automatically disambiguate the sense of these connectives. In this work only three specific connectors are analyzed, the corpus contains texts written in English, and machine learning is used. Nevertheless, we consider that some of the used features are interesting (mainly verbal tense), and we use them in our work. [15] also mention the difficulty of disambiguating discourse markers senses, especially when classifying rhetorical relations automatically. They propose a supervised machine learning method that uses several linguistic features to classify discourse relations in the absence of a cue phrase. They introduce the idea that tense and aspect offer clues about temporal relations and could influence the probabilities of different rhetorical relations. [10] use a Naive Bayes classifier to demonstrate that syntactic features improve performance in both discourse and non-discourse disambiguation tasks. In their experiments, they consider only the four top categories in the PDTB (Expansion, Comparison, Contingency and Temporal), obtaining a high accuracy in both experiments. Nevertheless, they do not offer a list and a linguistic analysis of the markers they use.

Some work on this subject is also found for other languages, such as German [16], [17] and Arabic [18]. For Spanish, there are few studies. We highlight the work of [19], who presents a proposal for detection and classification of Spanish discourse markers. Nevertheless, he mainly deals with the first type of ambiguity (sentence vs. discourse use of markers). In this study, punctuation (mainly the comma) is used to disambiguate discourse markers function. Later, [20] uses this work to create a system for detecting Spanish discourse markers automatically, but the main feature to identify different meanings is again punctuation.

After the revision of the state of the art, we can draw some conclusions: a) the disambiguation of discourse markers senses is a language-dependent task, since the lexical, syntactic and discourse features differ among languages; b) there is a research gap on this subject in Spanish; c) Spanish is a language with a high degree of syntactic complexity, and explicit Spanish discourse markers are more ambiguous than English markers, so the disambiguation task in Spanish is challenging; and d) to our knowledge, there are no studies carrying out a corpus analysis to detect the most frequent ambiguous markers in a language, or observing linguistic regularities in the different discourse relations they show.

3 Methodology

In the first stage, we use the database of Spanish discourse markers and RST relations proposed by [7] *to extract ambiguous discourse markers*. In other words, we extract discourse markers signalling more than one relation in the database. This database was created using the training corpus of the RST Spanish Treebank [21], which includes texts annotated with rhetorical relations (http://corpus.iingen.unam.mx/rst/). The corpus contains texts from nine specialized domains (Astrophysics, Earthquake Engineering, Economy, Law, Linguistics, Mathematics, Medicine, Psychology and Sexuality) and several genres (research articles, abstracts, sections of manuals and books, etc.). This variety of domains and genres guarantees that the results can be generalized. The corpus is divided into training corpus (183 texts) and test corpus (84 texts). It includes 52,746 words, 267 texts, 2,256 sentences and 3,349 discourse segments. The database mentioned above includes three types of markers:

1. Traditional discourse markers, such as *ya que* ("since").
2. Markers including lexical units, specifically, nouns and verbs, such as *metodología* ("methodology").
3. Markers including verbal structures, such as *para* ("to") + infinitive.

In our work, linguistic markers of Elaboration relations are not analyzed, since this is the most general and frequent relation in the language. We detect 31 markers indicating more than one relation in the training corpus. Over this list of ambiguous markers, two filters are applied: a) only the first and third types of discourse markers are analyzed, and b) only discourse markers with a frequency higher than the one in the corpus are taken into account. Thus, we obtain the 11 ambiguous discourse markers to be analyzed. Table 1 includes these markers and the marked relations (with their frequency in the training corpus indicated in brackets).

Table 1. Ambiguous discourse markers found in the corpus and analyzed in this work

Marker	Marked relations
pues ("since")	Cause (4), Justification (5)
ya que ("because")	Cause (2), Justification (3)
debido a ("due to")	Cause (7), Justification (4)
mientras ("while")	Contrast (11), Circumstance (2)
después ("after")	Sequence (3), Circumstance (4)
cuando ("when")	Condition (5), Circumstance (22)
y ("and")	Contrast (3), List (11)
o ("or")	Disjunction (6), Contrast (3)
al ("when") + infinitive	Cause (2), Circumstance (10)
comma + *lo que* ("which")	Interpretation (5), Result (6)
gerund verbal form	Concession (1), Condition (2), Result (7), Means (8), Circumstance (16)

In the second stage, *the discourse contexts of these ambiguous markers are extracted*, by using the RST_extract tool [22], which offers to the user text passages corresponding to discourse relations. We consider contexts to be: a) two EDUs (Nucleus and Satellite) in nucleus-satellite relations, and b) several EDUs (Nuclei) in multinuclear relations.

In the third stage, the contexts are analyzed automatically by using Freeling syntactic parser [23], available at: http://nlp.lsi.upc.edu/freeling/. Then, the contexts of each marker are analyzed manually, in order to find linguistic regularities in the contexts of each RST relation. These regularities are used *to develop rules capable of disambiguating the discourse markers senses*, that is, detecting the discourse relation they are marking in a specific context. The features that we analyze are:

- Verbal tense and mode (such as present vs. past, or indicative vs. subjunctive).
- Verbal lexical units (such as "to use", "to consider", etc.).
- Affirmative vs. negative verbal form.
- Position of the marker (such as at the beginning of the EDU).
- Combination of markers (such as "and + while").
- Subjects of the related EDUs.
- Punctuation (such as the comma).

In the fourth stage, *the developed rules are evaluated*, using the test corpus of the RST Spanish Treebank.

4 Analysis and Results

After analyzing the contexts of ambiguous discourse markers and detecting regularities, the disambiguation rules are designed and a template is created for each discourse marker (see Tables 2-10). In this corpus analysis, we observe different regularities, which are explained in this section.

I) The markers *pues* ("then", "since") and *ya que* ("because", "since") can express the relation of Justification or Cause. To justify an idea, speakers commonly use several arguments or related statements; therefore, sentences including a relation of Justification tend to contain several EDUs (usually with various discourse markers). On the contrary, to express the relation of Cause, speakers usually offer a fact first and then the cause of this fact directly, so the sentence includes only two EDUs (see Table 2). For example:

[*Los estudiantes adultos de origen chino, coreano y japonés tienen problemas para pronunciar los fonemas líquidos del españo*l]NUCLEUS [***ya que*** *en su lengua hay un solo fonema para estos sonidos.*]SATELLITE_CAUSE

[The adult students of Chinese, Korean and Japanese origin have problems to pronounce the liquid phonemes of the Spanish] [**since** in their language there is a single phoneme for these sounds.]

Table 2. Rule template for the markers *pues* ("then", "since") and *ya que* ("because", "since")

Discourse marker	Disambiguation rule
pues ("since") *ya que* ("because") ----------	IF 2 EDUs are related by the discourse marker *pues* ("then", "since") OR *ya que* ("because", "since") AND the 2 EDUs are included in a sentence consisting of only 2 EDUs
Marked relations	THEN relation = Cause
Cause	ELSE IF the 2 EDUs are included in a sentence consisting of more than 2 EDUs
Justification	THEN relation = Justification

In the case of the marker *debido a* ("due to"), it has not been possible to design a disambiguation rule. The regularities detected in the contexts of the markers *pues* and *ya que* have not been observed in the contexts of this marker, which can also express

Justification or Cause. Due to the lack of examples (two cases of Justification and three cases of Cause) we do not have enough information available, and more cases would be necessary in order to elaborate an adequate rule.

II) The marker *mientras* ("while") can signal the relation of Contrast or Circumstance. On the one hand, when making a Contrast between two elements, something is being argued or compared. On the other hand, the relation of Circumstance only offers some information or data (see Table 3). For example:

[***Mientras*** *se preparan dichas herramientas,*]SATELLITE_CIRCUMSTANCE [*habremos de trabajar sobre la modelización de los términos técnicos.*]NUCLEUS

[**While** these tools are prepared,] [we will have to work on the modelization of the technical terms.]

Table 3. Rule template for the marker *mientras* ("while")

Discourse marker	**Disambiguation rule**
mientras ("while") ---------- **Marked relations** Contrast Circumstance	@mientras = {*ya que* ("because, since"), *pues* ("then, since"), *por un/otro lado* ("on the one/other hand"), *por este/ese/aquel* ("on this/that case"), *en el primer/segundo/tercero/cuarto caso/lugar* ("in the first/second/third/fourth case/place"), *en este/ese/aquel caso* ("in this/that case")} IF 2 EDUs are related by the discourse marker *mientras* ("while") AND the marker is followed by the conjunction *que* ("that") OR the marker appears in combination with another discourse marker in @mientras THEN relation = Contrast ELSE IF the marker is not combined with *que* or another discourse marker in @mientras THEN relation = Circumstance

III) An EDU starting with the discourse marker *después* ("after") can be a part of a Sequence or indicate a Circumstance. On the one hand, if this marker appears in a segment constituting a single sentence, the relation should be Sequence, since the content is not offering a Circumstance of another segment. On the other hand, if the marker relates two segments in the same sentence, it could indicate a Circumstance (if the structure [después + *de* ("of") + infinitive] appears) or Sequence (if some other structure appears) (see Table 4). For example:

[*El virus se multiplica en las células y en la base de la lesión,*]NUCLEUS [*e infecta la neurona que los inerva (ganglio sacro).*]NUCLEUS [***Después*** *el virus volverá al punto inicial.*]NUCLEUS_SEQUENCE

[The virus is multiplied in the cells and in the base of the injury,] [and infects the neuron that innervates them (sacred ganglion).] [**Afterwards** the virus will return to the initial point.]

Table4. Rule template for the marker *después* ("after")

Discourse marker	**Disambiguation rule**
después ("after") ---------- **Marked relations** Sequence Circumstance	IF 2 EDUs are related by the discourse marker *después* ("after") AND the 2 EDUs are included in different sentences THEN relation = Sequence ELSE IF the 2 EDUs are included in the same sentence AND after *después* the preposition *de* ("of") appears, followed by an infinitive verbal form THEN relation = Circumstance ELSE IF the 2 EDUs are included in the same sentence AND the discourse marker is not followed by the structure *después* + *de* + infinitive verbal form THEN relation = Sequence

IV) The discourse marker *cuando* ("when") can indicate the relation of Circumstance or Condition. To determine which one of these two senses is correct, it is necessary to analyze the verbal tense and/or mode of the two EDUs that the marker relates. For example, if the main verb of the EDU containing the marker is a subjunctive verbal form, and the main verb of the other EDU is a present or future verbal form, the relation should be Condition; however, if the main verbs of the two EDUs are past forms, the relation should be Circumstance (see Table 5). For example:

[***Cuando*** *entramos a la sala de exhibición*]SATELLITE_CIRCUMSTANCE [*el susto fue inmenso.*]NUCLEUS

[**When** we enter in the exhibition room] [the fright was immense.]

Table 5. Rule template for the marker *cuando* ("when")

Discourse marker	Disambiguation rule
cuando ("when")	IF 2 EDUs are related by the discourse marker *cuando* ("when")
----------	AND the main verb of the EDU including the marker is a past verbal form
Marked relations	AND the main verb of the EDU not including the marker is a past verbal form
Condition	THEN relation = Circumstance
Circumstance	ELSE IF the main verb of the EDU including the marker is a present verbal form
	AND the main verb of the EDU not including the marker is a gerund verbal form
	THEN relation = Circumstance
	ELSE IF the main verb of the EDU including the marker is a subjunctive verbal form
	AND the main verb of the EDU not including the marker is a present OR future verbal form
	THEN relation = Condition
	ELSE IF the main verb of the EDU including the marker is a reflexive present verbal form
	AND the main verb of the EDU not including the marker is a present verbal form
	THEN relation = Circumstance
	ELSE IF the main verb of EDU including the marker is a non-reflexive present verbal form
	AND the main verb of the EDU not including the marker is a present verbal form
	THEN relation = Condition

V) Usually, in Spanish, the marker *y* ("and") marks the end of a List. Nevertheless, if this marker appears combined with another negation marker, it can indicate Contrast (see Table 6). For example:

[*No vulnera el sistema constitucional ni en general el orden jurídico*]NUCLEUS [***y sí, en cambio****, asegura que los derechos de la persona sean mejor protegidos y garantizados.*]NUCLEUS_CONTRAST

[It does not interfere in the legal order, neither in general in the constitutional system] [**and, by contrast,** it guarantees that people rights are better protected and guaranteed.]

Table 6. Rule template for the marker *y* ("and")

Discourse marker	Disambiguation rule
y ("and")	@contrast = {*no* ("no"), *en cambio* ("on the other hand"), *por el contrario* ("by contrast") *otro/otros/otra/otros* ("another/other/others")}

Marked relations	IF 2 EDUs are related by the discourse marker *y* ("and")
Contrast	AND *y* is combined with another discourse marker in @contrast
List	THEN relation = Contrast
	ELSE IF *y* is not combined with another discourse marker in @contrast
	THEN relation = List

VI) The marker *o* ("or") can signal the relation of Disjunction or Contrast. In the first case, the related EDUs have the same subjects while, in the second case, the subjects are different (see Table 7). For example:

[*¿Son términos todos los que lo parecen*]NUCLEUS **[*o abundan las creaciones léxicas sensacionalistas y efímeras?*]**NUCLEUS_CONTRAST

[Are all those that seem it terms] **[or** do the sensationalist and ephemeral lexical creations abound?]

Table 7. Rule template for the marker *o* ("or")

Discourse marker	Disambiguation rule
o ("or")	IF 2 EDUs are related by the discourse marker *o* ("or")
----------	AND the 2 EDUs have the same subject
Marked relations	THEN relation = Disjunction
Disjunction	ELSE IF the 2 EDUs have not the same subject
Contrast	THEN relation = Contrast

VII) In Spanish the construction [*al* ("when", "as") + infinitive] can be used to indicate a Cause or a Circumstance discourse relation (see Table 8). In the corpus, negative cases indicate Cause. For example:

[***Al no contar*** *en Cuba con propias referencias acerca del desarrollo del lenguaje infantil,*]SATELLITE_CAUSE [*se realizó una investigación nacional descriptiva y transversal.*]NUCLEUS

[**As** Cuba **does not have** its own references about the development of the infantile language,] [a descriptive and transversal national research was carried out.]

Table 8. Rule template for the marker *al* ("when", "as") + infinitive

Discourse marker	Disambiguation rule
al ("when") +	IF 2 EDUs are related by the syntactic construction *al* ("when", "as") + infinitive
infinitive	AND the construction includes a negation
----------	THEN relation = Cause
Marked relations	ELSE IF the 2 EDUs does not include a negation
Cause	THEN relation = Circumstance
Circumstance	

VIII) If an EDU starts with the relative pronoun *lo que* ("which") preceded by a comma, it can express the relation of Result or Interpretation. In this case, in order to differentiate both senses, the verb included in the EDU is used, since, in general, speakers use different verbs to express an objective result or their interpretation about something (e.g. *causar* ["to cause"] vs. *suponer* ["to supposse"]). For example:

[*Durante la pubertad, los niveles elevados de estrógenos hacen que el epitelio vaginal se adelgace y que el contenido de glucógeno celular se incremente,*]NUCLEUS [*lo que* ***provoca*** *que el pH vaginal disminuya.*]SATELLITE_RESULT

[During the puberty, the high levels of estrogens make the vaginal epithelium lose weight and the contents of cellular glycogen be increased,] [which **causes** that the vaginal pH decreases.]

Table 9. Rule template for the marker *lo que* ("which") preceded by a comma

Discourse marker	Disambiguation rule
lo que ("which") preceded by a comma --- **Marked relations** Interpretation Result	@interpretation = {*permitir* ("to allow"), *poner de manifiesto* ("to show"), *suponer* ("to suppose"), *conllevar* ("to entail")} @result = {*agudizar* ("to aggravate"), *causar* ("to cause"), *complicar* ("to complicate"), *conducir a* ("to lead to"), *dar lugar* ("to give place to"), *generar* ("to generate"), *hacer que* ("to cause"), *llegar* ("to arrive"), *manifestarse* ("to appear"), *obtener* ("to obtain"), *ofrecer* ("to offer"), *propiciar* ("to favour"), *provocar* ("to cause"), *resultar* ("to result"), *ser utilizado* ("to be used")} IF 2 EDUs are related by the relative *lo que* ("which") preceded by a comma AND the main verb of the EDU containing the relative is included in @result THEN relation = Result ELSE IF the main verb of the EDU containing the relative is included in @interpretation THEN relation = Interpretation

IX) In Spanish, gerund verbal forms should be used only to indicate simultaneity. Nevertheless, probably due to the influence of English, Spanish speakers tend to use gerunds to indicate Result, Concession, Means or Circumstance discourse relations. For example:

[*El objetivo de este trabajo es analizar los efectos de la política monetaria en el producto y los precios en la economía mexicana*]NUCLEUS [***utilizando*** *diversas técnicas econométricas.*]SATELLITE_MEANS

[The goal of this work is to analyze the effects of the currency policy in the product and the prices in the Mexican economy] [**using** different econometric techniques.]

Table 10. Rule template for the marker gerund verbal form

Discourse marker	Disambiguation rule
gerund verbal form --- **Marked relations** Concession Condition Result Means Circumstance	@result = {*agudizar* ("to aggravate"), *causar* ("to cause"), *complicar* ("to complicate"), *conducir a* ("to lead to"), *dar lugar* ("to give place to"), *generar* ("to generate"), *hacer que* ("to cause"), *llegar* ("to arrive"), *manifestarse* ("to appear"), *obtener* ("to obtain"), *ofrecer* ("to offer"), *propiciar* ("to favour"), *provocar* ("to cause"), *resultar* ("to result"), *ser utilizado* ("to be used")} @means = {*advertir* ("to advise"), *aplicar* ("to apply"), *aprovechar* ("to benefit"), *basarse* ("to be based on"), *comparar* ("to compare"), *controlar* ("to control"), *emplear* ("to use"), *esquematizar* ("to outline"), *estudiar* ("to study"), *hacer uso* ("to use"), *incluir* ("to include"), *incorporar* ("to incorporate"), *indagar* ("to investigate"), *plantear* ("to lay out"), *seguir* ("to continue"), *seleccionar* ("to select"), *tomar como base* ("to take as a base"), *tomar en cuenta* ("to take into account"), *trabajar* ("to work"), *usar* ("to use"), *utilizar* ("to use")} IF 2 EDUs are related by a gerund verbal form AND the gerund is preceded by the marker *aun* ("even") THEN relation = Concession ELSE IF the gerund is included in the EDU placed in the first position of the sentence THEN relation = Condition ELSE IF the gerund is not included in the EDU placed in the first position of the sentence AND the gerund is a verb included in @result THEN relation = Result AND the gerund is a verb included in the @means THEN relation = Means AND the gerund is not a verb included in @result OR @means THEN relation = Circumstance

Our disambiguation approach includes eight rules. These rules take into account all the features analyzed, except the punctuation feature. Unlike [19], we do not find that the comma offers relevant information in order to disambiguate discourse senses of Spanish markers. We consider that the reason is that, in Spanish, there are many cases in which the use of the comma is optional. Maybe this feature can help to differentiate between sentential and discourse uses of markers, but it is not useful to differentiate between their different discourse meanings.

5 Evaluation

In order to evaluate the performance of our approach, we use the test corpus of the RST Spanish Treebank, which constitutes a gold standard for Spanish. This corpus includes 84 texts, from the Mathematics, Psychology and Sexuality domains. Using again the RST_extract tool, we extract from this test corpus contexts with the following characteristics: a) the context includes an ambiguous discourse marker of our database; b) the context corresponds to one RST relation that can be expressed by that discourse marker, and c) a disambiguation rule has been created for this marker. 61 contexts are obtained.

Then, the disambiguation rules are applied to each context, in order to detect the RST relation that the contexts include. When applying the disambiguation rules, we assume that the EDUs related by the marker are previously detected. We obtain them by using the discourse segmenter DiSeg [24], available at: http://daniel.iut.univ-metz.fr/DiSeg/WebDiSeg/.

We calculate the number of contexts including ambiguous discourse markers that have been disambiguated correctly, obtaining an accuracy of 60.65%. Since there is no system developing this task in Spanish, we cannot compare our results to the results obtained with other approaches. Therefore, we create a baseline, following the methodology of the state of the art [13], [18]: the baseline offers the most frequent relation showed by the marker (in our case, in the training corpus of the RST Spanish Treebank). The baseline obtains an accuracy of 49.18%. For this reason, we consider that the results obtained in our experiments are representative, and constitute the first attempt towards the disambiguation of discourse markers senses in Spanish.

After a qualitative evaluation, we observe that the rules including lists of semantically related verbs (basically, the rules included in Tables 9 and 10) are useful, but they would have better performance if they included more verbs. In this study, we only include in the lists (@result, @interpretation and @means) the verbs found in the training corpus. In the test corpus, some contexts including different but semantically related verbs are detected. For example, *originar* (“to origin”), *traer* (“to bring along”) and *tener como consecuencia* (“to have as a consequence”) are semantically related to the verbs of the list @result; and the verbs *partir* (“to start from”), *iniciar* (“to begin”) and *abordar* (“to deal with”) are related to the verbs of the list @means. In addition, for some markers (such as *pues* and *ya que*), a few contexts are retrieved from the test corpus, so it is difficult to assess the performance of the corresponding rules.

6 Conclusions

In this paper, a symbolic approach to detect and solve the ambiguity of discourse markers in Spanish texts is presented. Specifically, we deal with discourse sense ambiguity, i.e. with markers that can signal more than one rhetorical relation (in this work, RST relations). The proposal is mainly based on syntactic and lexical features, and not on punctuation, as it has been done until now for Spanish. The performance of the approach is better than the baseline created following the methodology of the state of the art.

Although the results are encouraging, we are conscious that there is room for improvement. Specifically, as a future work, we will evaluate each rule individually, not the approach as a whole. Regarding the lack of contexts for the evaluation of some rules, [25] states that there are two possible strategies: a) to leave the corpus as it is, with few or no examples of some cases (but the problem will be the lack of training examples for machine learning systems), or b) to add low-frequency examples artificially in order to "enrich" the corpus (but the problem will be the distortion of the native frequency distribution and perhaps the confusion of machine learning systems). In the future, we plan to follow the second option, that is, to compile a specific corpus including contexts with ambiguous discourse markers, annotate it manually and then re-evaluate the problematic rules.

In addition, we plan to integrate semantic verbal information in the rules, to solve the problem detected in the qualitative evaluation, as mentioned in section 5. We will use lexical databases, such as EuroWordNet (http://www.illc.uva.nl/EuroWordNet).

Finally, in the future, our disambiguation approach will be integrated in a discourse parser for Spanish, and several related applications will be developed (automatic summarization and information extraction, among others). Also, we would like to combine our symbolic approach with machine learning methods, in order to examine the performance of a hybrid disambiguation system.

References

1. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3), 243-281 (1988)
2. Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. Computational Linguistics 26(3), 395-448 (2000)
3. Soricut, R., Marcu, D.: Sentence Level Discourse Parsing using Syntactic and Lexical Information. In: Proceedings of the 2003 Conference of NAACL-HLT, pp. 149-156 (2003)
4. Subba, R., Di Eugenio, B.: An effective discourse parser that uses rich linguistic information. In: Proceedings of the 2009 Conference of HLT-ACL, pp. 566-574 (2009)
5. Sumita, K., Ono, K., Chino, T., Ukita, T., Amano, S.: A discourse structure analyzer for Japanese text. In: Proceedings of the International Conference on Fifth Generation Computer Systems, pp.1133-1140 (1992)
6. Pardo, T.A.S., Nunes, M.G.V.: On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. Journal of Theoretical and Applied Computing 15(2), 43-64 (2008)
7. da Cunha, I., San Juan, E., Torres-Moreno, J-M., Cabré, M.T., Sierra, G.: A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-

sentence Discourse Segments in Spanish. In: Gelbukh, A. (ed). Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science (LNCS). CICLing 2012, Part I, 7181, pp. 462-474 (2012)
8. Maziero, E., Pardo, T.A.S., da Cunha, I., Torres-Moreno, J-M., SanJuan, E.: DiZer 2.0 - An Adaptable On-line Discourse Parser. In: Proceedings of the III RST Meeting (8 th Brazilian Symposium in Information and Human Language Technology) (2011)
9. Taboada, T.: Discourse markers as signals (or not) of rhetorical relations. Journal of Pragmatics 38, 567-592 (2006)
10. Pitler, E., Nenkova, A.: Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 13-16 (2009)
11. van Dijk, T.A.: Texto y contexto (Semántica y pragmática del discurso). Madrid: Cátedra (1984)
12. Hirschberg, J., Litman, D.J.: Empirical studies on the disambiguation of cue phrases. Computational Linguistics 19(3), 501–530 (1993)
13. Miltsakaki, E.; Dinesh, N.; Prasad, R.; Joshi, A.; Webber, B.: Experiments on sense annotations and sense disambiguation of discourse connectives. In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005) (2005)
14. Miltsakaki, E., Prasad, R., Joshi, A., Webber, B.:. The Penn Discourse Treebank. In 4th International Conference on Language Rescourses and Evaluation (LREC 2004), 2004.
15. Sporleder, C., Lascarides, A.: Exploiting Linguistic Cues to Classify Rhetorical Relations. In Proceedings of Recent Advances in Natural Language Processing (2005)
16. Bayerl, P.S.: Disambiguierung deutschsprachiger Diskursmarker: Eine Pilot-Studie. Linguistik Online, 18 (2004)
17. Versley, Y.: Multilabel Tagging of Discourse Relations in Ambiguous Temporal Connectives. In: Proceedings de la 8th International Conference on Recent Advances in Natural Language Processing, pp.154-161 (2011)
18. Al-saif, A., Markert, K.: Modelling Discourse Relations for Arabic. In: 2011 Conference on Empirical Methods in Natural Language Processing (2011)
19. Prada, J.J.: Marcadores del discurso en español. Análisis y representación. Master Thesis. Uruguay: Facultad de Ingeniería, Universidad de la República (2001)
20. Koza, W.A.: Detección automática de marcadores discursivos del español una aplicación con xfst. Philologica Urcitana. Revista de iniciación a la investigación en Filología 7, 59-74 (2012)
21. da Cunha, I., Torres-Moreno, J-M., Sierra, G.: On the development of the RST Spanish Treebank. In: Proceedings of the Fifth Law Workshop (ACL 2011), pp. 1-10 (2011)
22. da Cunha, I., Torres-Moreno, J-M., Sierra, G., Cabrera-Diego, L-A., Castro-Rolón, B-G.; Juan-Miguel Rolland-Bartilotti The RST Spanish Treebank On-line Interface. In: Angelova, G. et al. (eds.). *Proceedings of Recent Advances in Natural Language Processing*. pp. 698-703 (2011)
23. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3. Syntactic and semantic services in an open-source NLP library. In: N. Calzolari et al. (ed.). Proceedings of the Conference LREC 2006, pp. 48-55 (2006)
24. da Cunha, I., SanJuan, E., Torres-Moreno, J-M., Lloberes, M., Castellón, I.: DiSeg 1.0: The First System for Spanish Discourse Segmentation. Expert Systems with Applications 39(2), 1671-1678 (2012)
25. Hovy, E.: Annotation. A Tutorial. Presented at the 48th Annual Meeting of the Association for Computational Linguistics (2010)