DiSeg: Un segmentador discursivo automático para el español

DiSeg: An Automatic Discourse Segmenter for Spanish

Iria da Cunha
IULA, Universitat Pompeu Fabra
Laboratoire Informatique d'Avignon
C/ Roc Boronat 138, 08018, Barcelona
iria.dacunha@upf.edu

Eric SanJuan
Laboratoire Informatique d'Avignon
339, chemin des Meinajariés, Agroparc,
BP1228, 84911, Avignon
eric.sanjuan@univ.avignon.fr

Juan-Manuel Torres-Moreno
Laboratoire Informatique d'Avignon
339, chemin des Meinajariés, Agroparc,
BP1228, 84911, Avignon
juan-manuel.torres@univ.avignon.fr

Marina Lloberas
Universitat de Barcelona
C/ Gran Via de les Corts Catalanes 585,
08007, Barcelona
marina.lloberes@ub.edu

Irene Castellón
Universitat de Barcelona
C/ Gran Via de les Corts Catalanes 585,
08007, Barcelona
icastellon@ub.edu

Resumen: Hoy en día el análisis discursivo automático es un tema de investigación relevante. Sin embargo, no existen analizadores del discurso para textos en español. El primer paso para desarrollar esta herramienta es la segmentación discursiva. En este artículo presentamos DiSeg, el primer segmentador discursivo para el español que utiliza el marco de la *Rhetorical Structure Theory* (Mann y Thompson, 1988) y se basa en reglas léxicas y sintácticas. Describimos el sistema y evaluamos sus resultados con un corpus *gold standard*, obteniendo resultados prometedores.
Palabras clave: Análisis discursivo, segmentación discursiva, *Rhetorical Structure Theory*

Abstract: Nowadays discourse parsing is a very prominent research topic. However, there is not a discourse parser for Spanish texts. The first stage in order to develop this tool is discourse segmentation. In this work, we present DiSeg, the first discourse segmenter for Spanish that uses the framework of the Rhetorical Structure Theory (Mann and Thompson, 1988) and is based on lexical and syntactic rules. We describe the system and we evaluate its performance with a gold standard corpus, obtaining promising results.
Keywords: Discourse Parsing, Discourse Segmentation, Rhetorical Structure Theory

1 Introducción

Hoy en día el análisis automático del discurso es un tema de investigación relevante, ya que es útil para diversas aplicaciones, como el resumen automático, la generación de texto, la extracción de información, la traducción automática, etc. Existen analizadores discursivos para el inglés (Marcu, 2000a, 2000b), el japonés (Sumita et al., 1992) y el portugués de Brasil (Pardo et al., 2004; Pardo y Nunes, 2008). La mayoría emplean el marco de la *Rhetorical Structure Theory* (RST) (Mann y Thompson, 1988). Sin embargo, no existen analizadores discursivos para el español. El primer paso para desarrollar esta herramienta es realizar una segmentación discursiva automática, para posteriormente poder

determinar las relaciones discursivas existentes entre dichos elementos. Tal y como Tofiloski et al. (2009: 77) afirman[1]: "La segmentación discursiva es el proceso de descomponer el discurso en unidades del discurso elementales (EDUs), que pueden ser oraciones o cláusulas en una oración compleja, y a partir de las cuales se contruyen los árboles discursivos". Actualmente existen algunos segmentadores discursivos disponibles, para el inglés (Soricut y Marcu, 2003; Tofiloski et al., 2009[2]) y para el portugués de Brasil (Mazeiro et al., 2007[3]). No obstante, no conocemos ningún segmentador discursivo para el español.

En este trabajo presentamos DiSeg, el primer segmentador discursivo para el español, desarrollado en el marco de la RST. DiSeg formará parte de un analizador discursivo automático para el español que estamos desarrollando. Además, como herramienta individual, este segmentador puede ser útil para tareas que requieren anotación discursiva humana, ya que permitiría a los anotadores realizar sus análisis partiendo de una misma segmentación discursiva automática. También puede usarse esta herramienta como recurso para optimizar sistemas de resumen automático, permitiendo generar *abstracts* o resúmenes comprimidos (es decir, resúmenes que contienen no solo oraciones completas, sino también segmentos de oraciones). Además puede ser útil para aplicaciones de pregunta-respuesta o de extracción de información.

En este artículo describimos el sistema, basado principalmente en reglas léxicas y sintácticas que insertan fronteras discursivas en el interior de las oraciones. Evaluamos los resultados del sistema comparándolos con un corpus segmentado manualmente a modo de *gold standard*, ya que, al no existir otros segmentadores discursivos para el español, no podemos comparar nuestros resultados con los de otros sistemas similares. Sin embargo, como veremos más adelante, empleamos tres sistemas *baseline* en la evaluación. El sistema obtiene buenos y prometedores resultados, aunque pueden mejorarse algunos aspectos concretos.

En el apartado 2 presentamos la metodología del trabajo. En el apartado 3 explicamos la implementación del sistema. En el apartado 4 mostramos los experimentos y la evaluación. En el apartado 5 presentamos los resultados obtenidos. En el apartado 6 ofrecemos las conclusiones y el trabajo futuro.

2 Metodología

El primer paso de nuestra investigación fue decidir el marco teórico del que partir. Como ya hemos comentado en el apartado anterior, empleamos la RST. Como Taboada y Mann (2005) explican, esta teoría se ha empleado para investigar en diversos temas tanto teóricos como prácticos, como generación de texto (véase Hovy, 1993; Dale et al., 1992; O'Donnell et al., 2001, entre otros), resumen automático (véase Marcu, 2000a; Radev, 2000; Pardo y Rino, 2002, entre otros), traducción automática (véase Ghorbel et al., 2001; Marcu et al., 2000, entre otros), etc. Aunque en una gran cantidad de aplicaciones lingüísticas se emplean técnicas superficiales de tratamiento del texto con el objetivo de tratar grandes cantidades de datos, en nuestro trabajo empleamos la RST con el propósito de lograr un análisis lingüístico más profundo. Por ejemplo, en las investigaciones sobre resumen automático, normalmente se desarrollan sistemas por extracción, es decir, resumidores que incluyen oraciones completas extraídas del texto original. Usar estrategias basadas en la RST permitiría la eliminación de ciertos fragmentos del interior de las oraciones, obteniendo resúmenes más adecuados. Con respecto a la traducción automática, la tendencia actual es emplear técnicas estadísticas. Los resultados obtenidos pueden mejorarse con el análisis discursivo de la RST, al establecer paralelismos entre la estructura discursiva del texto original y del texto traducido.

La Figura 1 muestra un ejemplo de árbol discursivo de la RST. El árbol incluye dos relaciones núcleo-satélite (Concesión y Resultado) y una relación multinuclear con dos núcleos (Lista).

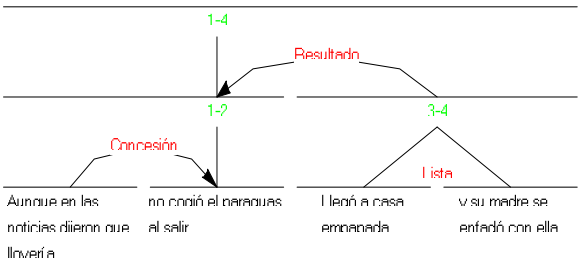

Figura 1: Ejemplo de árbol discursivo de la RST

[1] La traducción al español es nuestra.
[2] http://www.sfu.ca/~mtaboada/research/SLSeg.html
[3] http://www.nilc.icmc.usp.br/~erick/segmenter/

Una vez decidido el marco teórico, delimitamos la noción de EDU en nuestro trabajo. Usamos la noción de EDU empleada en el manual de Carslon y Marcu (2001), pero con algunas diferencias, similares a las incluidas en Tofiloski et al. (2009) y en da Cunha e Iruskieta (en prensa). El objetivo de estas diferencias es poder distinguir claramente el nivel sintáctico y el nivel discursivo. En nuestro trabajo, consideramos que una EDU debe incluir un verbo (es decir, debe constituir una oración o una cláusula) y debe mostrar, en sentido estricto, una relación discursiva o retórica[4] (en muchas ocasiones evidenciada por un conector discursivo). Por ejemplo, la oración 1a se dividiría en dos EDUs, mientras que la oración 1b constituiría una única EDU:

> 1a. [El hospital es conocido por sus tratamientos innovadores para enfermedades infecciosas raras,]$_{EDU1}$ [pero también tiene éxito en la cura de pacientes con enfermedades genéticas.]$_{EDU2}$
> 1b. [El hospital es conocido por sus tratamientos innovadores para enfermedades raras, además de las enfermedades genéticas.]$_{EDU1}$

Asimismo, las cláusulas de sujeto o de objeto no se considerarían EDUs. Por ejemplo, la oración 2 sería una única EDU:

> 2. [La enfermera les explicó que el servicio de urgencias del hospital donde ella trabajaba era muy eficaz y rápido.]$_{EDU1}$

Una vez determinada la noción de EDU, desarrollamos un conjunto de reglas de segmentación discursiva basadas en rasgos sintácticos y léxicos. Estas reglas están basadas, entre otros elementos, en:

1. Marcadores discursivos como "mientras", "aunque" o "es decir", que normalmente evidencian relaciones de Contraste, Concesión y Reformulación, respectivamente. Concretamente, empleamos el conjunto de marcadores discursivos incluidos en Alonso (2005).
2. Conjunciones, como, por ejemplo, "y" o "pero".
3. Adverbios, como "de todas maneras".
4. Formas verbales, como gerundios, verbos finitos, etc.
5. Signos de puntuación, como paréntesis o guiones.

A continuación implementamos las reglas de segmentación discursiva.

De cara a la evaluación del sistema, realizamos un corpus a modo de *gold standard*. Esta decisión se tomó, como ya adelantamos en en la introducción, debido a la carencia actual de segmentadores discursivos para el español. Uno de los mejores métodos para evaluar los resultados de un sistema es compararlos con los resultados de otros sistemas similares. Sin embargo, DiSeg constituye el primer segmentador discursivo para el español, por lo que no podemos usar otros sistemas para su evaluación.

Una vez implementado el sistema, evaluamos los resultados de DiSeg, empleando el *gold standard* desarrollado y las medidas de precisión, cobertura y *F-Score*. Además consideramos tres sistemas *baseline* y un sistema simplificado llamado DiSeg-base.

3 Implementación

La implementación de DiSeg incluye diversas etapas:

1. Segmentación oracional y *POS tagging*. En esta etapa empleamos Freeling (Asterias et al., 2006).
2. *Shallow parsing* con Freeling. Realizamos algunas modificaciones en la gramática de este *shallow parser*, principalmente recategorizaciones de algunos elementos en marcadores discursivos. El uso de este recurso *open-source* nos ha permitido realizar estas modificaciones, en las cuales se basan muchas de las reglas de segmentación discursiva que incluye DiSeg.
3. Transformación de la salida de Freeling a formato xml.
4. Aplicación de las reglas de segmentación discursiva:
   4.1. Detección de las fronteras entre segmentos (DiSeg-base). Esta detección integra dos autómatas simples basados en las siguientes

[4] En nuestro trabajo empleamos los adjetivos "discursivo" y "retórico" como sinónimos.

etiquetas: *ger, verb, vaux, forma_ger, ger_pas, coord, conj_subord, disc_mk* and *grup_sp_inf*. Además de estas etiquetas, los únicos marcadores que se usan en esta etapa son un signo de puntuación (la coma) y dos unidades léxicas: "que" y "para".

4.2. Definición de las EDUs (DiSeg). Si se aplicasen todas las fronteras detectadas en la etapa anterior, podrían llegar a generarse EDUs con ausencia de verbos. Para evitar esta situación, en esta etapa se analiza el texto de derecha a izquierda y, así, únicamente se consideran las fronteras como tal si existe un verbo en el segmento resultante antes y después de cada frontera.

Para las etapas 2, 3 y 4 usamos programas en Perl y Twig.

La Figura 2 refleja la arquitectura de DiSeg.

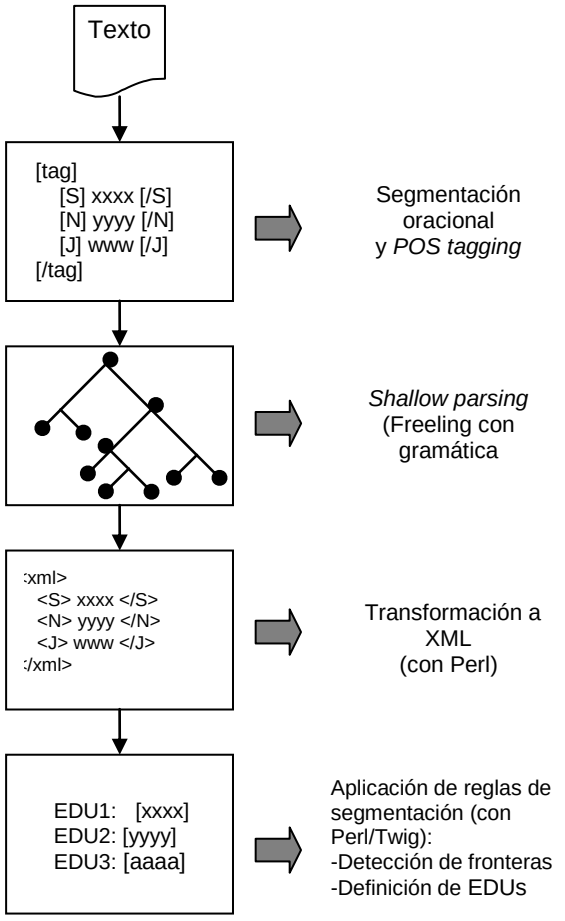

Figura 2: Arquitectura de DiSeg

La Tabla 1 incluye un fragmento del *gold standard* y la segmentación realizada por DiSeg. El *gold standard* completo puede consultarse en línea en: http://daniel.iut.univ-metz.fr/DiSeg/.

Texto original
Con el fin de predecir la tasa esperable de ganglios centinela en nuestra población, hemos analizado la tasa de invasión axilar en los últimos 400 casos de cáncer de mama pT1 operados por nosotros, utilizando la técnica clásica de linfadenectomía axilar completa. De los 400 tumores 336 (84.0%) fueron carcinomas ductales infiltrantes NOS, 32 (8.0%) carcinomas lobulillares, 22 carcinomas tubulares puros (5.5%), y los 10 restantes correspondieron a otras variedades histológicas menos frecuentes. A la hora de realizar el estudio del ganglio centinela en cánceres de mama T1 en nuestra población, cabe esperar globalmente la detección de un ganglio positivo en al menos una de cada cuatro pacientes.
Texto segmentado
<rst> **** Con el fin de predecir la tasa esperable de ganglios centinela en nuestra población, **** **** hemos analizado la tasa de invasión axilar en los últimos 400 casos de cáncer de mama pT1 operados por nosotros, **** **** utilizando la técnica clásica de linfadenectomía axilar completa. **** **** De los 400 tumores 336 (84.0%) fueron carcinomas ductales infiltrantes NOS, 32 (8.0%) carcinomas lobulillares, 22 carcinomas tubulares puros (5.5%), **** **** y los 10 restantes correspondieron a otras variedades histológicas menos frecuentes. **** **** A la hora de realizar el estudio del ganglio centinela en cánceres de mama T1 en nuestra población, **** **** cabe esperar globalmente la detección de un ganglio positivo en al menos una de cada cuatro pacientes. **** **</rst>**

Tabla 1: Ejemplo de segmentación de DiSeg

El Anexo 1muestra la captura de pantalla de la demo web del sistema, que puede utilizarse en línea en: http://daniel.iut.univ-metz.fr/~iula/WebDiSeg/index.php. El sistema también podrá descargarse en breve para que los usuarios puedan integrarlo en sus aplicaciones.

4 Experimentos y evaluación

El corpus *gold standard* para la evaluación incluye resúmenes de artículos médicos de investigación segmentados manualmente. Estos resúmenes se extrajeron de la revista médica en línea Gaceta Médica de Bilbao[5] y fueron recopilados por da Cunha e Iruskieta (en prensa). El corpus incluye 169 oraciones. La media de oraciones por texto es 8,45. El texto más largo contiene 21 oraciones y el más corto 3. El corpus contiene 3981 palabras. La media por texto es de 199 palabras. El texto más largo contiene 474 palabras y el más corto 59. El corpus incluye 203 EDUs. La media es de 10,15 EDUs por texto (el máximo es 28 EDUs y el mínimo 3 EDUs). En la Tabla 2 se resumen estas estadísticas, que son similares a las del corpus *gold standard* empleado por Tofiloski et al. (2009) para desarrollar su segmentador discursivo para el inglés, que obtuvo excelentes resultados.

	Total	**Texto mayor**	**Texto menor**	**Media**
Nº oraciones	169	21	3	8,45
Nº palabras	3981	474	59	199
Nº EDUs	203	28	3	10,15

Tabla 2: Estadísticas del *gold standard*

Este corpus fue segmentado por uno de los autores de este artículo (siguiendo las indicaciones de nuestro proyecto). Otro lingüista, externo al proyecto, segmentó el corpus siguiendo las mismas indicaciones. Calculamos precisión y cobertura de esta segunda anotación con respecto a la primera. Ambas medidas fueron altas: la precisión fue 98,05 y la cobertura 99,03. Las diferencias se debieron a errores de este segundo anotador. Después de un pequeño debate, llegamos a un consenso y, una vez resueltos los errores, empleamos esta segmentación como *gold standard*.

Una de las contribuciones relevantes de este trabajo es precisamente ofrecer un *gold standard* accesible para que la comunidad científica pueda emplearlo para la evaluación de nuevas propuestas de analizadores o segmentadores discursivos para el español.

Una vez conformado el corpus, aplicamos DiSeg sobre estos textos. Para evaluar los resultados empleamos medidas de precisión, cobertura y *F-Score* para contar las fronteras detectadas correctamente. La precisión es el número de fronteras en concordancia con el *gold standard*, y la cobertura es el número total de fronteras correctas detectadas por DiSeg dividido por el número de fronteras totales incluidas en el *gold standard*. Como en el trabajo de Tofiloski et al. (2009), no contamos las fronteras de oraciones para no mejorar artificialmente los resultados.

Para esta evaluación, utilizamos tres segmentadores *baseline*. La Baseline_0 únicamente considera las oraciones como EDUs. Esta no es una *baseline* trivial ya que su precisión es del 100% por definición y cuatro textos del *gold standard* no incluyen otro tipo de EDUs. La Baseline_1 inserta fronteras antes de la etiqueta *coor* introducida por el *shallow parser* de Freeling. La Baseline_2 toma como fronteras las etiquetas *coor* y *conj_subord*, pero solamente considera como EDU el último segmento a la derecha de la oración. También tenemos en cuenta en la evaluación el sistema simplificado DiSeg-base, donde se consideran todas las fronteras insertadas en la etapa 4.1., incluso aunque algunas de las EDUs generadas no incluyan verbo. Para realizar una evaluación homogénea con DiSeg, no contamos las fronteras de las oraciones como segmentos en ninguno de los sistemas *baseline*.

5 Resultados

La Tabla 3 contiene los resultados de la evaluación. Los resultados muestran que el sistema DiSeg completo supera a DiSeg-base y a las otras tres *baselines*. Las diferencias de *F-Score* son estadísticamente significativas según el test de Student (0,05 entre DiSeg y 0,01 entre DiSeg y los tres sistema *baseline*; la Baseline_2, la más sofisticada, es la que ofrece los mejores resultados).

[5] http://www.gacetamedicabilbao.org/web/es/

	Precisión	Cobertura	*F-Score*
DiSeg	**71%**	**98%**	**80%**
DiSeg-base	70%	88%	74%
Baseline_2	68%	82%	72%
Baseline_1	33%	70%	39%
Baseline_0	100%	49%	62%

Tabla 3: Resultados de la evaluación

Estos resultados son muy similares a los obtenidos por el segmentador discursivo para el inglés de Tofiloski et al. (2009): 93% de precisión, 74% de cobertura y 83% de *F-Score*. Por tanto, consideramos que los resultados de DiSeg son prometedores.

Después del análisis cuantitativo de los resultados, llevamos a cabo un análisis cualitativo para detectar los principales errores del sistema. Encontramos problemas relacionados con las reglas de segmentación discursiva y problemas respecto a Freeling. El principal problema de las reglas de segmentación se refiere a situaciones donde el elemento "que" aparece al mismo tiempo que la conjunción "y". El ejemplo 3 ilustra la situación: el ejemplo 3a muestra la segmentación de DiSeg y el 3b refleja la segmentación correcta.

> 3a. [El perfil del usuario sería el de un varón (51,4%) de mediana edad (43,2 años) que consulta por patología traumática (50,5%)]$_{EDU1}$ [y procede de la comarca sanitaria cercana al hospital.]$_{EDU2}$

> 3b. [El perfil del usuario sería el de un varón (51,4%) de mediana edad (43,2 años) que consulta por patología traumática (50,5%) y procede de la comarca sanitaria cercana al hospital.]$_{EDU1}$

Una de las reglas de segmentación de DiSeg indica que el relativo "que" no debe considerarse como frontera de segmento. Sin embargo, otra de las reglas indica que si se encuentra una conjunción coordinada (como "y") y más adelante un verbo, dicha conjunción constituye una posible marca de segmentación. Así, DiSeg no segmenta antes de "que", pero sí segmenta justo antes de la "y", ya que detecta el verbo "procede" antes del final de la oración. Hemos encontrado diversos casos con una problemática similar.

Además, hemos detectado dos errores derivados de una segmentación oracional incorrecta de Freeling. El ejemplo 4 muestra uno de ellos: el ejemplo 4a refleja la segmentación de DiSeg y el ejemplo 4b la segmentación correcta.

> 4a. [No encontramos cambios en la medición del ángulo astrágalo-calcáneo en AP. Realizamos una descripción de nuestra serie y una discusión acerca de la técnica y de la indicación actual de la cirugía en esta patología.]$_{EDU1}$

> 4b. [No encontramos cambios en la medición del ángulo astrágalo-calcáneo en AP.]$_{EDU1}$ [Realizamos una descripción de nuestra serie y una discusión acerca de la técnica y de la indicación actual de la cirugía en esta patología.]$_{EDU2}$

El módulo de segmentación oracional no segmenta correctamente estas dos oraciones, probablemente porque considera "AP." como una abreviación y no detecta el inicio de la segunda oración. Este problema provoca un error en la segmentación discursiva de Diseg. De todas maneras, este tipo de errores puede solucionarse fácilmente, ya que Freeling permite cambiar y añadir información a su diccionario.

6 Conclusiones

En este trabajo hemos desarrollado DiSeg, el primer segmentador discursivo para el español, basado en reglas léxicas y sintácticas y en el marco de la RST. Consideramos que esta investigación constituye un paso importante en la investigación sobre análisis discursivo automático en español, ya que existen muy pocos trabajos sobre este tema para esta lengua.

Hemos evaluado los resultados de DiSeg, midiendo precisión, cobertura y *F-Score*. Lo hemos comparado con un *gold standard* que hemos desarrollado. Los resultados son positivos si los comparamos con los de los sistemas *baseline*. Además, son similares a los obtenidos por Tofiloski et al. (2009). Consideramos que el *gold standard* que hemos creado es una buena contribución para animar a otros investigadores a continuar trabajando en esta línea.

Como trabajo futuro planeamos solucionar los problemas detectados en cuanto a la precisión, usando más reglas simbólicas y/o aprendizaje automático. Además, prevemos aumentar el tamaño del corpus y aplicar DiSeg sobre otro corpus en español que contenga textos de ámbito general de la Wikipedia.

También prevemos aplicar DiSeg a tareas de Procesamiento del Lenguaje Natural (PLN), como resumen automático, para llevar a cabo una evaluación extrínseca.

El objetivo final del proyecto es desarrollar el primer analizador discursivo automático para el español basado en la RST, que integraremos en una plataforma abierta, fácilmente adaptable a otras lenguas latinas compatibles con Freeling. Para desarrollar este analizador seguiremos dos estrategias principales: por un lado, utilizaremos patrones léxicos que evidencien relaciones (en la línea de Pardo y Nunes, 2008) y, por otro, aplicaremos aprendizaje automático (en la línea de Afantenos et al., 2010).

Agradecimientos

Parte de este trabajo ha sido financiado mediante una ayuda de movilidad posdoctoral otorgada por el Ministerio de Ciencia e Innovación de España (Programa Nacional de Movilidad de Recursos Humanos de Investigación; Plan Nacional de Investigación Científica, Desarrollo e Innovación 2008-2011) a Iria da Cunha.

Bibliografía

Afantenos, S., P. Denis, P. Muller y, L. Danlos (2010). "Learning Recursive Segments for Discourse Parsing". En *Proceedings of the Seventh conference on International Language Resources and Evaluation*.

Alonso, L. 2005. "Representing discourse for automatic text summarization via shallow NLP techniques". Tesis doctoral. Barcelona: Universitat de Barcelona.

Atserias, J., B. Casas, E. Comelles, M. González, Ll. Padró, y M. Padró. 2006. "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library". En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, ELRA.

Carlson, L. y D. Marcu. 2001. *Discourse Tagging Reference Manual*. ISI Technical Report ISITR-545. Los Ángeles: University of Southern California.

da Cunha, I. y M. Iruskieta (en prensa). "Comparing rhetorical structures of different languages: The influence of translation strategies". *Discourse Studies* 12(5).

Dale, R., E. Hovy, D. Rösner, y O. Stock (Eds.). 1992. *Aspects of Automated Natural Language Generation*. Berlín: Springer.

Ghorbel, H., A. Ballim, y G. Coray. 2001. "ROSETTA: Rhetorical and Semantic Environment for Text Alignment". En P. Rayson, A. Wilson, A. M. McEnery, A. Hardie, y S. Khoja (Eds.). *Proceedings of Corpus Linguistics 2001*. 224-233. Lancaster, UK.

Hovy, E. 1993. "Automated discourse generation using discourse structure relations". *Artificial Intelligence*, 63. 341-385.

Mann, W.C. y S.A. Thompson. 1988. "Rhetorical structure theory: Toward a functional theory of text organization". *Text*, 8(3): 243-281.

Marcu, D. 2000a. *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts: Institute of Technology.

Marcu, D. 2000b. "The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach". *Computational Linguistics*, 26(3): 395-448.

Marcu, D., L. Carlson, y M. Watanabe. 2000. "The automatic translation of discourse structures". En *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL'00). Vol. 1. 9-17. Seattle, Washington.

Mazeiro, E.G. y T.A.S. Pardo. 2009. "Metodologia de avaliação automática de estruturas retóricas". En *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* (STIL). São Carlos, Brasil: Universidade de São Paulo.

Erick G. Mazeiro, Thiago A.S. Pardo and Maria das Graças V. Nunes. 2007. "Identificação automática de segmentos discursivos: o uso do parser PALAVRAS". Série de Relatórios do Núcleo Interinstitucional de Lingüística

Computacional (NILC). São Carlos, São Paulo.

O'Donnell, M. 2000. "RSTTOOL 2.4 – A Markup Tool for Rhetorical Structure Theory". En *Proceedings of the International Natural Language Generation Conference*. 253-256.

O'Donnell, M., C. Mellish, J. Oberlander, y A. Knott. 2001. "ILEX: An architecture for a dynamic Hypertext generation system". *Natural Language Engineering*, 7. 225-250.

Pardo, T.A.S. y L.H.M. Rino. 2002. "DMSumm: Review and assessment". En *Proceedings of Advances in Natural Language Processing, Third International Conference (*PorTAL 2002). 263-274. Faro, Portugal: Springer.

Pardo, T.A.S., M.G.V. Nunes, y L.H.M. Rino. 2004. "DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese". *Lecture Notes in Artificial Intelligence*, 3171: 224-234.

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

Radev, D. 2000. "A common theory of information fusion from multiple text sources. Step one: Cross document structure". En L. Dybkjær, K. Hasida and D. Traum (Eds.). *Proceedings of 1st SIGdial Workshop on Discourse and Dialogue*. 74-83. Hong-Kong.

Soricut, R. y D. Marcu. 2003. "Sentence Level Discourse Parsing Using Syntactic and Lexical Information". En *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 149-156. Edmonton, Canada.

Sumita, K., K. Ono, T. Chino, T. Ukita, y S. Amano. 1992. "A discourse structure analyzer for Japonese text". En *Proceedings of the International Conference on Fifth Generation Computer Systems*. 1133-1140.

Taboada, M. Y W.C. Mann. 2005. "Applications of rhetorical structure theory". *Discourse Studies*, 8(4): 567-588.

Tofiloski, M., J. Brooke y M. Taboada. 2009. "A Syntactic and Lexical-Based Discourse Segmenter". En *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*. Singapur.

A Anexo 1: Captura de pantalla de la demo web de DiSeg