Lexicón Computacional de Marcadores del Discurso *

Laura Alonso
CLiC, Department of General Linguistics
Universitat de Barcelona
lalonso@lingua.fil.ub.es

Irene Castellón
Department of General Linguistics
Universitat de Barcelona
castel@lingua.fil.ub.es

Lluís Padró Cirera
TALP Research Center
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
padro@lsi.upc.es

Resumen: Presentamos la construcción de un lexicón computacional de marcadores del discurso orientado a resumen automático. Hemos realizado un análisis de una primera versión del mismo mediante métodos empíricos e implementación en un sistema de resumen. Este análisis ha dirigido una nueva configuración del lexicón, más adecuada descriptivamente y proceduralmente. Asimismo, hemos identificado algunas características generales de los marcadores como operadores discursivos.
Palabras clave: marcadores del discurso, resumen automático, lexicón

Abstract: We present the construction of a computational lexicon of discourse markers oriented to text summarization. We have carried out an analysis of a first version of this lexicon with empirical methods and by implementation in a text summarization system. This analysis has directed a new configuration of the lexicon, more adequate both descriptively and procedurally. Moreover, we have identified some general features of discourse markers as discourse operators.
Keywords: discourse markers, automated text summarisation, lexicon

1 *Motivación*

La caracterización de cierta estructura del discurso se ha abordado frecuentemente utilizando marcas superficiales que permitan su representación de forma parcial. Estas marcas, que llamamos marcadores del discurso (MDs) (Schiffrin, 1987), resultan muy útiles para tareas de PLN complejas de amplia cobertura, ya que aportan información muy rica sobre la estructura discursiva, con un bajo coste de procesamiento.

Sin embargo, un obstáculo para su uso generalizado en PLN es que la comunidad científica no ha alcanzado el consenso respecto a su delimitación y caracterización. Esta falta de consenso se debe, por un lado, a la preeminencia de las aproximaciones de tipo deductivo, con un sesgo importante por una teoría subyacente, y por otro, a la subordinación de la mayor parte de caracterizaciones a una tarea computacional concreta, lo que suele conllevar soluciones *ad hoc*.

En este artículo presentamos la construcción, análisis y reformulación de un lexicón computacional de MDs. Aunque este recurso está orientado para su uso en tareas de resumen automático, hemos llevado a cabo un análisis exhaustivo que pretende dar cuenta de características generales de los MDs en cuanto a operadores de procesamiento del discurso, realizando una descomposición analítica de sus elementos de significado fundamentales, de forma que su caracterización resulte lo suficientemente flexible para ser utilizada en diversas aplicaciones. Hemos apoyado este análisis en técnicas empíricas para minimizar un posible sesgo teórico.

En el resto de esta sección presentamos nuestro marco de trabajo. En la Sección 2 planteamos una delimitación de la unidad de trabajo, el MD. En la Sección 3 describimos la configuración inicial del lexicón, junto con los problemas surgidos al implementarlo, y en la Sección 4 exponemos las mejoras propues-

* Este trabajo se ha realizado gracias a una beca pre-doctoral asociada al proyecto X-Tract, PB98-1226 del Ministerio de Educación y Cultura. Parte de esta investigación se inscribe en el marco de los proyectos HERMES (TIC2000-0335-C03-02) y PETRA (TIC2000-1735-C02-02).

tas, para concluir en la Sección 5.

1.1 Marco de Trabajo

Los MDs se utilizan para obtener una estructura del discurso útil para un buen número de aplicaciones de PLN, como por ejemplo Resolución de Co-Referencia (Cristea et al., 1999), Gestión de Diálogos (Di Eugenio, Moore, y Paolucci, 1997; Kim, Glass, y Evens, 2000) o Resumen Automático (Ono, Sumita, y Miike, 1994; Marcu, 1997a), pero también en tareas básicas de procesamiento del discurso, como la identificación de relaciones de relevancia y coherencia (Edmunson, 1969), segmentación discursiva de los textos (Alonso y Castellón, 2001) e incluso para derivar una cierta estructura del discurso (Marcu, 1997a).

Expresiones como *porque*, *a pesar de* o *en ese caso* se encuentran entre los MDs más utilizados, porque resultan muy informativos de la estructura del discurso. Estas marcas pueden ser tratadas de forma satisfactoria con técnicas de PLN superficiales. También los signos de puntuación, algunas estructuras sintácticas y, en general, diversas claves textuales superficiales pueden caracterizar la estructura del discurso de modo análogo.

Aquí presentamos un lexicón computacional de MDs del español. Este lexicón está implementado en un sistema de resumen genérico y de extracción, que puede funcionar autónomamente o en colaboración con otras técnicas o tipos de información. El sistema está compuesto por dos módulos que utilizan el lexicón (véase Figura 1):

- el **segmentador** detecta de las unidades discursivas básicas (segmentos) y
- el **interpretador** pondera cada segmento según su relevancia discursiva.

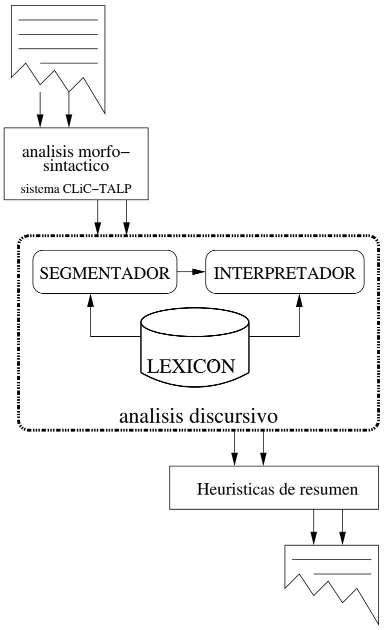

Figura 1: Sistema de ayuda al resumen automático mediante MDs

2 *Delimitación de la unidad de trabajo*

Las investigaciones sobre MDs realizadas para el español abordan el concepto de MD desde el punto de vista descriptivo, mayoritariamente en estudios de análisis del discurso, gramatical o lexicográfico, como queda reflejado en el compendio de Martín Zorraquino y Montolío (1998). La definición que sintetiza este tipo de aproximaciones la encontramos en Martín Zorraquino y Portolés (1999):

> *unidad lingüística invariable que no realiza una función sintáctica en la predicación oracional, es decir, que realiza una función discursiva.*

Sin embargo, la aproximación de estos trabajos a los MDs resulta insuficiente para su uso en PLN, tanto en extensión como en intensión. Por un lado, no disponemos aún de una relación exhaustiva de las partículas discursivas del español[1]. Además, el tipo de unidades tratadas es demasiado restringido para dotar de una cobertura aceptable a un sistema automático de análisis discursivo de ámbito general. Por otro lado, la tradición lingüística en lengua española ofrece una caracterización de los MDs que, pese a su adecuación descriptiva, escapa al nivel de formalidad necesario para la implementación.

Entre los trabajos sobre MDs específicamente orientados a PLN, cabe destacar el de Dale y Knott (1995), que proponen mecanismos formales para la detección y sistematización de estas unidades. Knott (1996) aplica estos mecanismos para obtener y caracterizar un conjunto de unos 200 MDs de la lengua inglesa. Sobre este trabajo, y asumiendo las premisas básicas de la Rhetorical Structure Theory, RST (Mann y Thompson, 1988), Marcu (1997b)

[1]Hay que señalar que se encuentran en marcha al menos un proyecto para desarrollar un recurso lexicográfico de estas unidades.

desarrolla un sistema de análisis de la estructura retórica para el inglés basado en la información discursiva que obtiene de un conjunto de 400 MDs. Sin embargo, ninguno de estos dos trabajos soluciona la creación de un listado extenso y no controvertido de MDs para uso computacional, o cómo abordar la creación de estos recursos para otras lenguas.

En Alonso, Castellón, y Padró (2002) presentamos X-TRACTOR, una herramienta para la adquisición automática de MDs mediante técnicas de bootstraping. Se parte de un pequeño conjunto de MDs prototípicos, que son traducidos a características definitorias, características que dirigen la búsqueda de nuevos MDs sobre corpus de gran tamaño. Al ser identificados por un proceso automático, los nuevos MDs estarán menos sesgados que aquellos procedentes de observaciones humanas, tanto por prejuicios teóricos como por limitaciones de extensión[2]. En resultados preliminares hemos observado que los MDs obtenidos por este método resultan muy cercanos a nuestro concepto intuitivo de MD. Así pues, esta puede ser una buena aproximación a la delimitación del concepto de MD por extensión.

Sin embargo, una aproximación por extensión parece insuficiente como definición autónoma del concepto. Por esta razón hemos tomado la siguiente definición como punto de referencia:

> *es* **MD** *toda unidad textual, léxica, sintáctica o de puntuación, que marca la presencia de un segmento discursivo (Alonso y Castellón, 2001) o de una relación discursiva entre segmentos.*

Como se desprende de esta definición, prácticamente cualquier elemento textual puede ser considerado MD. El criterio para determinar si un elemento textual es un MD viene determinado por su función en los textos y por nuestra capacidad de identificar y utilizar esa función mediante nuestras herramientas de PLN. Para ello partimos de análisis del discurso con jueces humanos, pero también nos ayudamos de técnicas más empíricas, como la herramienta de extracción automática de MDs X-TRACTOR.

[2]La extensión de corpus que puede cubrir una herramienta automática es mayor que la de un juez humano, aunque la profundidad puede ser menor.

3 *Lexicón de MDs*

Realizamos una primera versión del lexicón de MDs para su uso en un sistema de ayuda al resumen automático, tomando como punto de partida diversas fuentes de información:

- el conjunto de MDs propuestos por Knott (1996) y Marcu (1997c)
- las listas de palabras invariables del formario del analizador morfológico MACO (Arévalo et al., 2001)
- un estudio de corpus

En esta primera versión se describieron 577 MDs básicos[3], y su caracterización se basó en los siguientes parámetros (véase Tabla 1):

- **delimitador**: un MD puede ser un delimitador de segmentos *fuerte*, si es una marca no ambigua de límite de segmento discursivo, *débil*, si los segmentos que introduce no son discursivamente independientes, o bien pueden no marcar límites de segmento.
- **tipo sintáctico**: distinguimos *adverbial*, *anafórico*, *conjunción*, *no personal*, *preposicional* y *subordinante*.
- **tipo retórico**: siguiendo la Rhetorical Structure Theory, (RST), consideramos *satelizadores* a aquellos MDs que identifican un segmento retóricamente subordinado y *nucleizadores* a los que identifican un segmento más central en el discurso. En relaciones simétricas entre segmentos, consideramos *encadenador* al MD que señala una relación de co-dependencia, *organizador* al que explicita una estructura organizada, y *conector* al que no evidencia ninguna relación.
- **dirección**: Dependiendo del segmento al cual está más estrechamente ligado, distinguimos MDs orientados a la *derecha*, a la *izquierda*, *ambiguos* a derecha o izquierda, *bi-direccionales*, o por *inclusión*, si se encuentran en el interior del segmento.
- **contenido retórico**: Cada MD recibe un contenido retórico, basado en las relaciones de la RST estándar (Mann y Thompson, 1988).

[3]Las expansiones morfológicas de estos 577 MDs dan un resultado de 784 formas

MD	delimitador	tipo sintáctico	tipo retórico	dirección	contenido
además	no aplicable	adverbial	satelizador	inclusión	refuerzo
a pesar de	fuerte	preposicional	satelizador	derecha	concesión
así que	débil	subordinante	encadenador	derecha	consequencia
dado que	débil	subordinante	satelizador	derecha	capacitación

Tabla 1: Muestra de la primera versión del lexicón de MDs

3.1 Implementación y problemas

Implementamos este lexicón en un sistema de ayuda al resumen automático, presentado en la Figura 1. El módulo de segmentación discursiva explota la información de delimitación, tipo sintáctico y dirección de los MDs, mientras que tipo retórico y contenido son utilizados en la fase de interpretación.

Esta primera implementación dió buenos resultados para el nivel oracional, es decir, para la detección e interpretación de segmentos entre signos de puntuación fuertes. Sin embargo, al abordar niveles de mayor alcance, como el párrafo y el texto, se han puesto en evidencia carencias e inadecuaciones de la caracterización de los MDs.

En primer lugar, observamos que el conjunto de relaciones retóricas de la RST estándar no respondía satisfactoriamente a las necesidades de la tarea de resumen según nuestras capacidades de PLN, sino que era o bien demasiado específico o bien demasiado general. Por otro lado, no se recogía la estructura discursiva lineal de los textos. Por ejemplo, el párrafo que se ve en la Figura 2, cuando un análisis más adecuado de este mismo texto sería el que encontramos en la Figura 3, que incluye información lineal-argumentativa.

La implementación de este recurso también ha mostrado inadecuaciones en la organización interna del lexicón, como el hecho de que ciertas características aportaban información redundante o que se trataban conjuntamente tipos de información heterogénea (descriptiva y procedural). Por estos motivos realizamos una exploración del lexicón que nos permitiera fundamentar una mejor configuración del mismo, tanto en el diseño como en el contenido.

4 Reformulación del lexicón de MDs

Nuestro principal objetivo al reconfigurar el lexicón fue mejorar su adecuación descriptiva, tanto en lo que se refiere a los MDs que contiene como a los atributos que los caracterizan. Por otro lado, quisimos mejorar el diseño del recurso, eliminando información redundante y dotándolo de una arquitectura modular, que facilite su portabilidad.

4.1 Mejora Estructural

Para analizar la organización estructural del lexicón, aplicamos técnicas clustering. Obtuvimos ocurrencias en corpus de los MDs, y describimos cada ocurrencia como un vector, mediante algunas características de su contexto de ocurrencia y los atributos asociados al MD en el lexicón. Clasificamos estos vectores mediante Klass+ (Gibert, 1997), una herramienta de clustering orientada a dominios poco estructurados. Esta herramienta creó clases con las ocurrencias de MDs más similares entre sí, y determinó las características que definían cada una de estas clases. De esta forma, observamos que las características de *dirección* y *tipo sintáctico* eran mútuamente redundantes, y que era más adecuado trabajar con tres grupos de comportamiento sintáctico: *extra-oracional*, *bidireccional* y *a la derecha*.

Asimismo, el análisis por clustering evidenció que las características *contenido retórico* y *tipo retórico* eran mútuamente redundantes dentro de un mismo grupo de comportamiento sintáctico. Creemos que eso es debido a que se trata de tipos heterogéneos de información: información descriptiva e información procedural. Siguiendo la propuesta de Knott (2000), hemos organizado la información procedural como instrucciones de procesamiento asociadas al lexicón, que ahora es puramente descriptivo. Cada MD se asocia, mediante sus atributos descriptivos, a una o más instrucciones de procesamiento básicas, que describen su comportamiento como señal de procesamiento del discurso.

4.2 Mejoras de contenido

Como hemos expuesto en la Sección 3.1, el conjunto de relaciones propuesto por la RST no resuelve de manera satisfactoria las necesidades de representación del discurso requeridas por la tarea de resumen automático, y

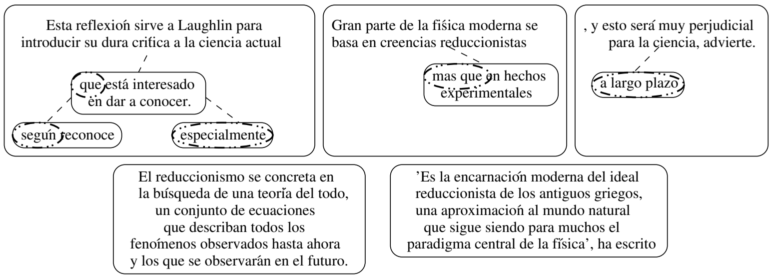

Figura 2: Análisis de un párrafo basado en relaciones retóricas de tipo jerárquico (líneas discontínuas) basadas en MDs (en círculos), que se establecen entre segmentos discursivos (en cajas).

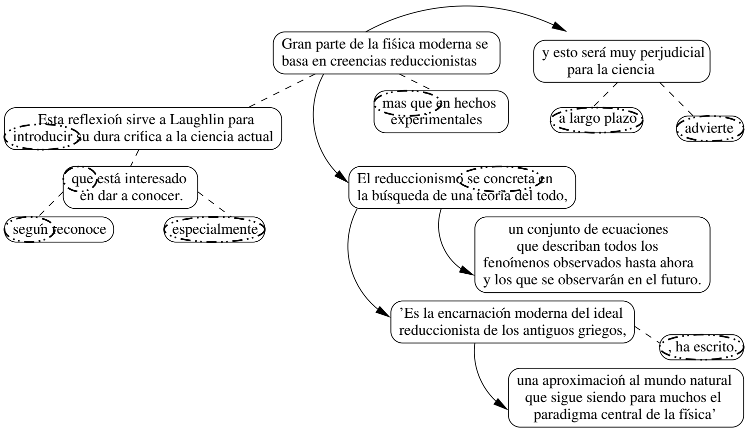

Figura 3: Análisis de un párrafo integrando relaciones retóricas, de tipo jerárquico (líneas discontínuas) con relaciones argumentativas, de tipo lineal (flechas). Las flechas a la derecha indican *progresión*, hacia abajo indican *elaboración*, y a la izquierda indican *revisión*.

tampoco responde a las características de las herramientas de PLN de qué disponemos, ya que propone distinciones que no podemos determinar con suficiente fiabilidad.

Además, a menudo se ha hecho notar que la coherencia textual no está determinada únicamente por la estructura jerárquica del discurso como la propuesta por la RST, sino también por la precedencia lineal (Grosz y Sidner, 1986). Habitualmente, la precedencia lineal se trata con la resolución de las referencias entre entidades o eventos. Sin embargo, la referencia a argumentos y las relaciones entre entidades o eventos son también un elemento importante en la configuración del discurso.

Para intentar capturar el aspecto jerárquico y lineal de la estructura del discurso, hemos establecido un conjunto de relaciones retóricas definidas por tres parámetros (véase Figura 4):

- relaciones **de materia**: se establecen entre entidades del discurso[4]. Hemos adoptado los tres grandes grupos de relaciones retóricas que identifican Asher y Lascarides (2002) (*figure-ground, similarity, causality*), distinguiendo subtipos cuando ha resultado necesario. Los hemos definido de la siguiente forma:

  – **contexto** (*figure-ground*): sitúan una entidad en un contexto determinado. Dentro de este grupo suele resultar útil para muchas tareas distinguir entre diversos tipos de circunstanciales (tiempo, lugar, etc.).
  – **causalidad** (*causality*): establecen una relación de causa entre dos entidades. En algunos casos es posible considerarlas como un subtipo de las relaciones de contexto.
  – **paralelismo** (*similarity*): estable-

[4]Tal como proponen Webber et al. (1999), consideramos los eventos también como entidades del discurso.

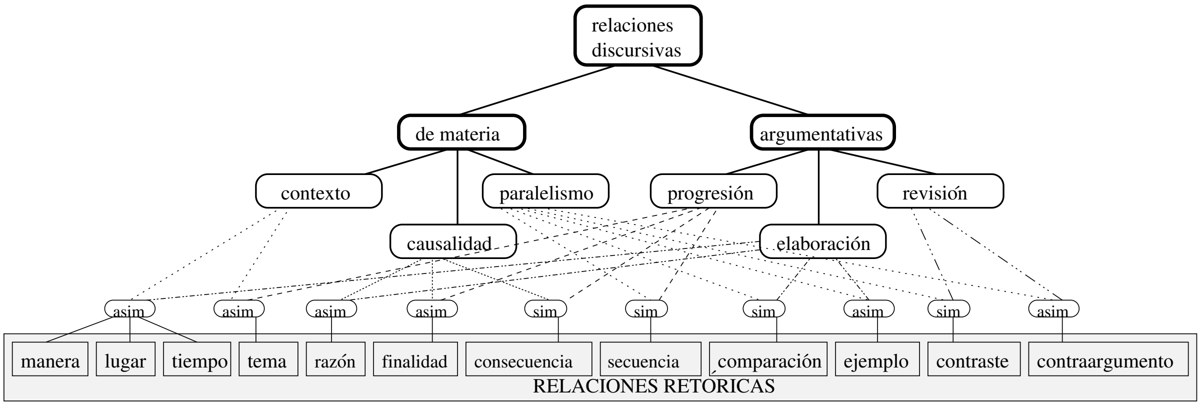

Figura 4: Conjunto de relaciones retóricas (en rectángulos) definidas en tres parámetros: relaciones de materia, relaciones argumentativas y configuración estructural del discurso.

cen una relación de equivalencia entre dos entidades, del tipo *A es (modalidad) B*.

- relaciones **argumentativas**: partimos de la idea de Anscombre y Ducrot (1983) de que todo texto tiene una orientación argumentativa, y que cada uno de sus elementos contribuye de alguna forma a esta orientación. A diferencia de las relaciones de materia, éstas se establecen entre fragmentos de texto, no entre entidades. Hemos distinguido tres tipos:
  - **progresión**: dado que la función primera de un discurso es desarrollar una línea argumentativa de forma lineal, consideramos que todo fragmento de texto contribuye a su progresión argumentativa a no ser que se explicite lo contrario.
  - **elaboración**: aquellos fragmentos de texto que no introducen información nueva o, si lo hacen, se trata de información sin una continuidad en el discurso posterior, sino que únicamente contribuye a restringir las inferencias posibles sobre la información que se ha dado hasta ese momento.
  - **revisión**: algunos fragmentos de texto suponen una revisión de la información contenida en otros fragmentos, anteriores o posteriores. Esta revisión supone un aumento en el coste de procesamiento del discurso, por lo cual esta relación es la más marcada de las argumentativas.
- configuración de la estructura jerárquica del discurso: las relaciones **simétricas** se establecen entre segmentos en el mismo nivel de la jerarquía, mientras que las relaciones **asimétricas** se establecen entre un nivel superior y otro inferior. Para distinguir entre estos tipos de relaciones recurrimos a la noción de frontera derecha propuesto por Polanyi (1988) y aplicado a la distinción de relaciones discursivas coordinantes o subordinantes por Asher y Vieu (2001).

De esta forma, cada relación retórica queda definida por sus valores en cada uno de los tres aspectos que hemos descrito. La organización en una tipología nos permite infraespecificar los marcadores más ambiguos, por ejemplo, asignándole la etiqueta *contexto* cuando no es posible determinar uno de los subtipos de este grupo de relaciones.

4.3 Configuración final del lexicón de MDs

Para hacer el lexicón más flexible, para que pueda ser utilizado en diversas tareas, se ha separado la información descriptiva, más independiente de tarea, de la procedural, específica de aplicación. Así, tenemos un módulo con los atributos descriptivos de cada MD, y otro con sus instrucciones de procesamiento asociadas. Sobre este último módulo se basan principalmente las herramientas de resumen automático que utilizan el lexicón.

El módulo descriptivo consta de los siguientes campos (véase Tabla 2):

- **peso léxico**: se obtiene de la longitud del MD (en palabras) y la frecuencia de las palabras que lo forman ($longitud \cdot \log(frecuencia)$), la cifra obtenida se traslada a una escala del 0 al 3. Esta característica contribuye a determinar la ambigüedad de algunos MDs en cuanto a su función discursiva u oracional, ya que, cuanto cuanto mayor es el peso léxico de un MD, menor es su ambigüedad como operador discursivo.
- **sintaxis**: a partir del estudio de clustering, hemos determinado tres tipos de comportamiento sintáctico: *a la derecha*, *bi-direccional*, *extra-oracional*.
- **contenido retórico**: hemos asociado a cada marcador una de las relaciones retóricas presentadas en la Sección 4.2.

Por lo que respecta al módulo de procesamiento, asociamos a cada MD ciertas instrucciones de procesamiento según su configuración de atributos. Podemos distinguir entre instrucciones para determinar el tipo de MD e instrucciones para el análisis del discurso. Las primeras sirven para identificar la función que realiza una ocurrencia determinada de un MD del lexicón mediante sus atributos y algunas características contextuales. Los atributos asociados a cada MD determinan qué instrucciones de exploración del contexto deben realizarse y cómo deben interpretarse para determinar las instrucciones de análisis del discurso.

La organización de las instrucciones por orden de especificidad nos permitirá establecer una estructura de tipos de MD en la cual se puedan aplicar relaciones de inferencia. Una estructura así también asegura la coherencia en la caracterización de los MDs, ya que los mecanismos de infraespecificación estarán claramente explicitados, y los MDs con un comportamiento discursivo similar, caracterizados por atributos también similares en el lexicón, compartirán todas o parte de sus instrucciones de procesamiento.

5 Conclusiones y Direcciones Futuras

Hemos presentado la descripción y análisis de un lexicón computacional de MDs mediante su implementación en un sistema de resumen automático y aplicando métodos empíricos. Hemos detectado algunas inadecuaciones de tipo descriptivo, como el hecho de que no se tenía en cuenta información de estructura lineal del discurso, y de tipo estructural, como la redundancia entre características descriptivas de los MDs o la no distinción entre informaciónes heterogéneas.

Hemos incorporado estas mejoras en una segunda versión de este recurso. Esta versión presenta una arquitectura modular que la hace más flexible que la anterior, por lo tanto, más fácil de integrar en sistemas de PLN distintos al de resumen automático presentado aquí. La aplicación efectiva de este recurso a otras tareas es una línea de trabajo futuro a explorar.

El trabajo presentado en este artículo apunta que una aproximación empírica puede contribuir significativamente a la delimitación del concepto de MD para tareas de PLN. Como hemos visto, la implementación y el uso de métodos empíricos resultan útiles para identificar y organizar las características definitorias de los MDs. Como trabajo futuro nos planteamos la ampliación de este recurso mediante la adquisición automática de nuevos MDs, la aplicación de este lexicón a distintas tareas de PLN, su adaptación a distintas lenguas y el uso de métodos empíricos para comprobar la adecuación de la herramienta, con el objetivo de construir un recurso de amplia cobertura y que contribuya a la delimitación del concepto de MD.

Bibliografía

Alonso, Laura y Irene Castellón. 2001. Towards a delimitation of discursive segment for natural language processing applications. En *First International Workshop on Semantics, Pragmatics and Rhetoric*, Donostia - San Sebastián, November.

Alonso, Laura, Irene Castellón, y Lluís Padró. 2002. X-tractor: A tool for extracting discourse markers. En *LREC 2002 workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Las Palmas.

Anscombre, J. C. y O. Ducrot. 1983. *L'argumentation dans la langue*. Mardaga.

Arévalo, Montse, Laura Alonso, Mariona Taulé, y M. Antònia Martí. 2001. Documentación sobre el analizador morfológico para el castellano (amcas). Informe

MD	peso léxico	sintaxis	contenido retórico
además	2	extra-oracional	elaboración
a pesar de	3	derecha	contra-argumento
así que	1	derecha	conscuencia
dado que	2	derecha	razón

Tabla 2: Muestra de la segunda versión del lexicón de MDs

Técnico X-Tract 01/01 Working Paper, CLiC, Universitat de Barcelona.

Asher, Nicholas y Alex Lascarides. 2002. *The Logic of Conversation*. Cambridge University Press.

Asher, Nicholas y Laure Vieu. 2001. Subordinating and coordinating discourse relations. En *Intl. Wkshp. on Semantic, Pragmatics and Rhetorics*, San Sebastián.

Cristea, Dan, Daniel Marcu, Nancy Ide, y Valentin Tablan. 1999. Discourse structure and co-reference: An empirical study. En *The ACL99 Workshop on Standards and Tools for Discourse Tagging*.

Dale, Robert y Alistair Knott. 1995. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Di Eugenio, Barbara, Johanna D. Moore, y Massimo Paolucci. 1997. Learning features that predict cue usage. En *ACL-EACL97*, Madrid, Spain.

Edmunson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264 – 285, April.

Gibert, Karina. 1997. The use of symbolic information in automation of statistical treatment of ill-structured domains. *Artificial Intelligence Communications*.

Grosz, Barbara y Candance Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204.

Kim, Jung Hee, Michael Glass, y Martha W. Evens. 2000. Learning use of discourse markers in tutorial dialogue for an intelligent tutoring system. En *COGSCI 2000*.

Knott, Alistair. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. tesis, University of Edinburgh, Edinburgh.

Knott, Alistair. 2000. An algorithmic framework for specifying the semantics of discourse relations. *Computational Intelligence*, 16(4):501–510.

Mann, William C. y Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organisation. *Text*, 3(8):234–281.

Marcu, Daniel. 1997a. From discourse structures to text summaries. En Mani y Maybury, editores, *Advances in Automatic Text Summarization*, páginas 82 – 88.

Marcu, Daniel. 1997b. The rhetorical parsing of natural language texts. En *ACL-97*, Madrid, Spain, 7 - 12 July.

Marcu, Daniel. 1997c. *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts*. Ph.D. tesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Martín Zorraquino, M. Antonia y Estrella Montolío, editores. 1998. *Los marcadores del discurso: teoría y análisis*. Arco Libros, Madrid.

Martín Zorraquino, M. Antonia y José Portolés. 1999. Los marcadores del discurso. En Ignacio Bosque y Violeta Demonte, editores, *Gramática Descriptiva de la Lengua Española*, volumen III. Espasa Calpe, Madrid, páginas 4051–4213.

Ono, K., K. Sumita, y S. Miike. 1994. Abstract generation based on rhetorical structure extraction. En *COLING-94*, páginas 344 – 348, Kyoto, Japan.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge University Press.

Webber, Bonnie, Alistair Knott, Madeleine Stone, y Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalised tag. En *ACl-99*.