IMPORTANCIA DEL MARCAJE DE LAS RELACIONES DISCURSIVAS PARA LA GENERACIÓN AUTOMÁTICA DE RESÚMENES[1]

IRIA DA CUNHA FANEGO
Instituto Universitário de Lingüística Aplicada (IULA)
Universidad Pompeu Fabra

1. INTRODUCCIÓN

Podemos definir un resumen (*abstract)* como una condensación de los conceptos principales del contenido del texto al que hace referencia (Burgos *et al*, 1994). El American National Standars Institute (ANSI) define el *abstract* en el ámbito científico como: "an abbreviated, accurate representation of the contents of a document, preferably prepared by its author(s) for publication with it "(Bhatia, 1993).

Al elaborar un resumen pueden seguirse diferentes criterios (no excluyentes entre ellos) en función de los cuales puede hacerse una clasificación a partir de tres factores: el documento a resumir (*input*), el resumen obtenido (*output*) y el propósito del mismo (*purpose*).

El *input* puede ser un único documento frente a varios, o textos que versen sobre un dominio específico frente a documentos del ámbito general, o también documentos monolingües frente a multilingües. En cuanto al *output*, debe considerarse si el resumen que se quiere obtener es una reestructuración coherente del texto o bien una extracción de los segmentos más relevantes del *input* (*abstract vs. extract*), un resumen fluido o no, o incluso si debe ser neutral o evaluativo. En relación con el propósito para el que se redactará el resumen, debemos tener en cuenta si queremos hacernos simplemente una idea general de los aspectos contenidos en el texto, o de si necesitamos una información más detallada (resumen indicativo vs. informativo). También el resumen será diferente dependiendo de si refleja el punto de vista del autor o si debe responder a alguna cuestión del usuario en concreto; y, ya por último, puede asumirse que el lector es un lego en la materia de la que trata el texto o, por el contrario, que tiene un conocimiento muy amplio, con lo cual el tipo de resumen también variará.

Este estudio se centrará en aportar datos en cuanto a la generación automática de resúmenes (*abstracts*), un problema complejo sobre el que se está trabajando desde diversas perspectivas, como veremos a continuación. La mayoría de las aportaciones en este campo no se centran en aspectos propiamente lingüísticos, sino estadísticos, por lo que los resultados aún dejan bastante que desear. El punto de vista que se defiende en este trabajo es que las relaciones discursivas que existen dentro de un texto desempeñan un papel fundamental para conseguir un buen resumen del mismo, pero veremos que, aun así, puede que su consideración no sea suficiente.

Al condensar las ideas más relevantes de un texto en otro de menor extensión que mantenga cohesión y coherencia, muchos de los sistemas actuales de generación automática de resúmenes se basan en aspectos cuantitativos de los textos, mediante técnicas estadísticas de recuperación de información (Berger *et al*, 2000; Brandow *et al*, 1994; Dunning, 1993). De todos modos, creemos que en este campo de investigación

[1] Este artículo se enmarca en un estudio más amplio que servirá de base a una tesis doctoral sobre generación automática de resúmenes dirigida por el doctor Leo Wanner (Universidad de Stuggart / Universidad Pompeu Fabra, Barcelona).

deben tratarse aspectos lingüísticos ya que, aunque los textos puedan verse desde esa perspectiva, no son entidades matemáticas.

Hovy (2003) divide los métodos actuales de generación automática de resúmenes en tres tipos: métodos superficiales, métodos de nivel medio y métodos profundos. En cuanto a los métodos superficiales se incluyen pistas dadas por frecuencias de palabras (Luhn 1959; Edmunson, 1969), posición de determinados fragmentos (Lin & Hovy, 1997), títulos (Luhn, 1959) o *cue phrases* (palabras clave) (Edmundson 1969; Teufel & Monees, 1999). Varios de estos métodos sirvieron para iniciar la investigación sobre el resumen automático. Por lo que se refiere a los métodos de nivel medio, se incluyen técnicas de reconocimiento de cadenas léxicas, elementos relacionados o correferencia (Barzilay *et al*, 1997; Boguraev *et al*, 1997); también hay métodos basados en la máxima de relevancia marginal (Goldstein and Carbonell, 1999). Finalmente, podría trabajarse con métodos profundos como la utilización de la estructura del discurso (Marcu, 1998; Ono *et al*, 1994). Es en este último aspecto en el que nos centraremos en este artículo**.**

Los sistemas avanzados de generación automática de resúmenes se basan en la idea de que un texto viene definido por su estructura interna y las relaciones discursivas que la forman (Marcu, 1997a). Estos sistemas toman como base teórica la *Rethorical Sructure Theory* (Mann & Thompson, 1988) para explicar cómo las relaciones discursivas de un texto muestran la estructura del mismo. La metodología utilizada para evidenciar la importancia de estas relaciones se basará en el marcaje de árboles de estructuras y en el de los marcadores discursivos. Estos marcadores son piezas léxicas con significación propia, formadas por uno o más morfemas (léxicos i/o gramaticales), que guían a los receptores de un texto en la descodificación del discurso en que se incluyen orientándolos hacia una conclusión determinada (Bach, 2001). Estas unidades juegan aquí un papel destacable ya que muchas veces son marcas que evidencian relaciones internas del texto. Dependiendo del tipo de relación discursiva marcada en el texto se seleccionarán o desecharán determinados fragmentos del mismo, todo ello orientado a una posterior aplicación de generación automática de resúmenes.

Esta idea es importante a la hora de resumir un documento, ya que es obvio que hay algunos fragmentos más relevantes que otros dependiendo de la información que aporten. Así, en estas relaciones, en ocasiones señaladas por marcadores discursivos que las unen, hay fragmentos que deberían incluirse en un resumen, mientras que otros podrían obviarse.

El objetivo de este trabajo es llevar a cabo el marcaje de las relaciones discursivas de un corpus de textos médicos para evaluar si el análisis de su estructura discursiva y de sus marcadores es suficiente para llegar a resumirlos automáticamente, o si para lograrlo deberíamos utilizar además otro tipo de información, como la que aportan la estructura informativa (o comunicativa) y sintáctica de los textos.

2. CORPUS DE ANÁLISIS

Los sistemas que intentan resumir textos del ámbito general pueden no dar buenos resultados, ya que las estructuras de estos textos pueden ser diferentes, así como también lo pueden ser los fenómenos discursivos e informativos que se producen en ellos, y el estilo de cada autor o género que se quiera analizar. Así, en un primer estadio de la investigación nos hemos centrado en analizar textos de un determinado dominio específico, que sí ofrezcan posibilidades reales de coherencia y exactitud. Aun así hay que tener en cuenta la posibilidad de que los datos y conclusiones futuras puedan ser extrapolables a otro tipo de géneros.

En este trabajo nos centraremos en analizar artículos científicos de medicina, y para ello hemos utilizado el Corpus Técnico del Instituto Universitario de Lingüística Aplicada (IULA).[2] Se ha seleccionado un subcorpus en español que se adapte a nuestros intereses de investigación, formado por artículos de la revista *Medicina Clínica*, única publicación semanal de contenido clínico que se edita en España y que constituye el máximo exponente de la calidad y pujanza de la medicina española. Son características fundamentales de esta publicación el rigor científico y metodológico de sus artículos y la actualidad de los temas. Hemos seleccionado artículos acompañados de sus correspondientes resúmenes o *abstracts* redactados por el mismo autor. Esto nos servirá en un futuro para realizar diversas pruebas y comprobar si las aproximaciones al resumen automático coinciden con el resumen del autor. Somos conscientes de que en muchas ocasiones el *abstract* inicial no se corresponde con el artículo que finalmente redacta el escritor, pero para evitar ambigüedades hemos realizado pruebas con jueces humanos para comprobar que efectivamente el resumen recoge las ideas fundamentales del documento. Esta correspondencia se debe tanto a calidad de la revista, como a la clara estructura textual de este tipo de artículos, de la que hablaremos a continuación.

3. LAS ESTRUCTURAS DE TEXTOS

Una vez aclarado el corpus utilizado y las motivaciones que llevan a ello, debe tenerse en cuenta que los textos son entidades que pueden ser vistas desde diversas perspectivas. De hecho, podemos encontrarnos con una estructura en cuanto a la organización de los textos, y con otra que se refiera a las relaciones discursivas que existen en su interior.

Tenemos un primer nivel de análisis formado por las divisiones textuales propias del género *artículo científico de medicina* (en concreto cuatro), y un segundo nivel, el de la estructura discursiva, que viene dado por la relaciones que se integran en cada una de las divisiones de la estructura textual, y que pueden aparecer evidenciadas por marcadores discursivos.

3.1. *Estructura textual (1er nivel)*

La estructura textual de los artículos científicos de medicina y de sus *abstracts* correspondientes viene claramente definida por ser normalmente fija.

Así Salager-Meyer (1991) lleva a cabo un estudio sobre la organización discursiva de los *abstracts* de artículos científicos en inglés y llega a la conclusión de que estos tienen una estructura retórica similar a la de sus correspondientes artículos, es decir, que pasan por cuatro *moves* para expresar el orden lógico de pensamiento científico: Introducción, Métodos, Resultados y Discusión (IMRD). Diversos investigadores españoles han realizado estudios sobre este tema pero orientándolos hacia nuestra lengua, y han llegado a conclusiones similares, incluso estableciendo apartados dentro de los *moves* (López Arroyo, 2002), o diferenciando entre resumen

[2] Este corpus recoge textos escritos en cinco lenguas diferentes (castellano, catalán, inglés, francés y alemán) y de cinco ámbitos especializados distintos, como son: derecho, economía, medioambiente, medicina e informática. Estos textos son clasificados por especialistas en cada materia, marcados con códigos SGML y posteriormente sometidos a una cadena de procesamiento. Una vez finalizado este proceso, los textos se incorporan a Bwananet, una interfaz que permite la consulta de este Corpus Técnico vía Internet (http://brangaene.upf.es/bwananet/index.htm).

informativo (IMRD) e indicativo, menos utilizado en castellano (Lorés, 2002). Por lo tanto, desde el punto de vista textual, la estructura de estos artículos suele ser la misma**.**

Esta información nos ayudará a determinar un primer nivel de análisis, el de la estructura textual. Una vez hayamos llegado a él, podrá pasarse al siguiente, el de la estructura discursiva, que hemos mencionado más arriba. Es en este punto donde deben considerarse las relaciones discursivas que se encuentran en cada uno de los apartados de la estructura textual.

3.2. *Estructura discursiva (2º nivel)*

Hobbs (1978, 1985) fue uno de los primeros estudiosos en describir de manera formal la estructura interna de los textos y en desarrollar un formalismo para representarlas. Como se ha mencionado más arriba, Mann & Thompson (1988) desarrollan a su vez la *Rethorical Structure Theory*, que se ha hecho muy popular en el análisis de textos. La RST es una teoría descriptiva de organización del texto muy útil para describirlo caracterizando su estructura a partir de las relaciones que mantienen entre sí los elementos del mismo (circunstancia, elaboración, motivación, evidencia, justificación, causa, propósito, antítesis, condición, entre otras). Se basa a su vez en una serie de afirmaciones, como la predominancia de estructuras con patrones de núcleo-satélites, la funcionalidad de la jerarquía y el rol comunicativo de la estructura del texto. Establecen un listado de relaciones internas del texto, en las que algunos de sus elementos (satélites) aportan ciertas informaciones acerca de la otra parte de la relación (núcleo), que es más esencial que la anterior. Estos satélites no serían comprensibles separados de su núcleo, y podrían ser fácilmente sustituibles.

Basándose en la RST, Marcu (1996) parte de la segmentación del texto en unidades mínimas y del ya mencionado conjunto de relaciones que pueden mantener entre ellas, para proporcionar una formalización de la estructura retórica arbórea usando la distinción entre el núcleo y los satélites que pertenecen a las relaciones discursivas. Ofrece un algoritmo basado en un conjunto de rasgos para la construcción de posibles árboles retóricos. Posteriormente, relaciona esto con el resumen automático y crea un prototipo de sistema que utiliza estos análisis retóricos para seleccionar las unidades textuales más relevantes del texto. Cuanto mayor sea la longitud del resumen que se quiera generar, más se alejaran los elementos relevantes de la raíz de la estructura arbórea (Marcu, 1997a). El autor deriva la estructura de los textos usando marcadores del discurso, y para este fin utiliza un sistema que posee un amplio conjunto de estas unidades y de las relaciones que pueden existir entre elementos, siempre ampliables según las necesidades del usuario (Marcu, 1997b). En trabajos posteriores Marcu (1998, 2000) sigue profundizando acerca de las estructuras de los textos, del *parsing* automático y del resumen automático.

Siguiendo la metodología de la RST hemos marcado manualmente las relaciones discursivas de nuestro corpus, utilizando como soporte la RSTtool de Marcu. A continuación, en la Fig.1, mostramos el resultado de una pequeña muestra en formato de árbol de relaciones. Se observa una relación de contraste entre los dos núcleos (1-4), es decir que se trata de una relación multinuclear. Además cada uno de ellos tiene satélites. El primer núcleo tiene un satélite que está formado por dos elementos de una lista (2-3), mientras que al segundo le sigue un satélite que deja ver una relación concesiva (4-5). Marcu (1997a) se basa en este tipo de estructuras discursivas para posteriormente delimitar los segmentos más relevantes del texto, desechando los que no aportan información esencial a la hora de resumir.

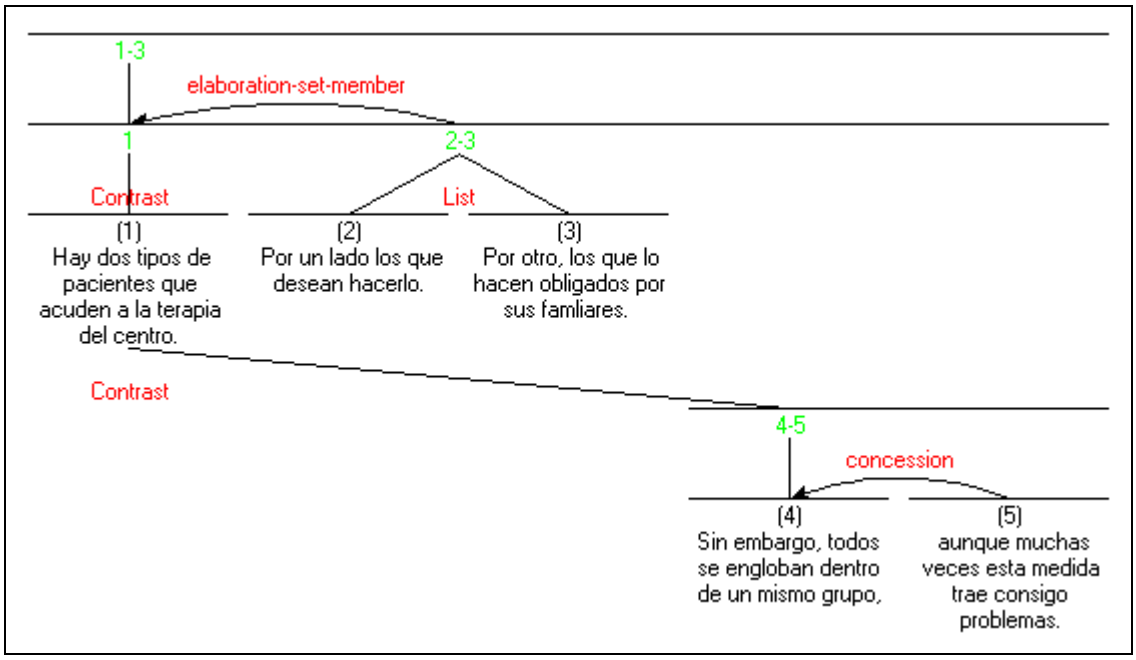

Fig.1. Fragmento de estructura arbórea de relaciones discursivas realizada con la RSTtool de Marcu como soporte.

Así, hay ciertos elementos de las relaciones discursivas que podríamos eliminar si queremos hacer un resumen, como, por ejemplo, los fragmentos evidenciados por conectores reformulativos. Veremos un caso en el siguiente fragmento:

> Ej. 1. "<1er elemento> <s> El 7% de los hombres y el 20% de las mujeres de la Comunidad Valenciana han sufrido al menos una depresión a partir de los 35 años de edad. </s> </1er elemento>
> <2º elemento> <s> <conector> *Es decir* </conector>, las mujeres de la Comunidad Valenciana tienen una mayor tendencia que los hombres a la depresión en edad adulta.</s> </2º elemento>"

En este ejemplo el marcador *es decir* introduce una reformulación parafrástica total (Bach, 2001), con lo cual el segundo elemento de la relación podría ser eliminado, ya que ofrece la misma información que el primero pero enunciado de otra forma. A la hora de confeccionar un resumen, sería innecesario incluir las dos oraciones.

Esta propuesta puede ser efectiva, pero la dificultad surge al intentar clasificar automáticamente las relaciones discursivas existentes en los textos. Para este fin debemos fijarnos en los marcadores discursivos que relacionan elementos, aunque estos no aparecerán en todas la relaciones. Se necesitaría un listado completo y exhaustivo que reflejase todos los marcadores del discurso para llegar a una aplicación computacional basada en ellos y, aunque existen en castellano varios listados de marcadores, todavía no hay un consenso general, ya que cada autor utiliza su propia definición y, en consecuencia, su propio listado. Se está trabajando mucho sobre este tema y hay estudios realmente interesantes, como el de Alonso y Castellón (2001) que en una fase inicial de su investigación establecen una primera aproximación al tratamiento computacional de las palabras clave de la estructura textual (estableciendo un subcorpus de 388 marcadores del discurso), e incluso desarrollan una herramienta de segmentación. Afirman también que "en dominios específicos: puede haber marcadores

también específicos, algunos marcadores de significado muy general pueden tener un significado más concreto, y se puede reducir la ambigüedad de los marcadores". Ésta será una de las líneas futuras de nuestro estudio, ya que nos estamos basando en el dominio restringido de la medicina.

Hay además otras investigaciones y tesis que también han tratado el tema de los marcadores del discurso orientados hacia aplicaciones informáticas. Uno de los ejemplos más representativos es el de Prada (2001), que hace una propuesta para el reconocimiento automático e interpretación estructural de estos marcadores. Presenta además una clasificación de los mismos, dividiéndolos en estructuradores, conectores, reformuladores, operadores argumentativos, y conversacionales. Se basa en la propuesta de Portolés (1998), uno de los estudiosos que más se ha centrado en trabajar los marcadores discursivos.

Por otro lado hay un proyecto de investigación vigente cuyo objetivo ha sido la redacción de un *Diccionario de Partículas Discursivas del Español (DPD)*, basado en las últimas investigaciones en España acerca de los marcadores del discurso. La motivación proviene de que "el español carece de un diccionario que describa el uso de sus partículas discursivas" (Briz, 2003).

4. DISCUSIÓN: PRUEBA DE DENSIDAD DE MARCADORES

Hemos explicado dos niveles de análisis que tienen suma importancia a la hora de elaborar un resumen. Pero somos conscientes de que hay dificultades. Por un lado, con referencia a la estructura textual (primer nivel), podría no cumplirse en algún texto concreto de medicina la máxima de la estructuración cuatripartita, por ejemplo. Sin un primer nivel de análisis podría suceder que el establecimiento de la estructura discursiva (segundo nivel) no fuese suficiente para lograr un resumen coherente.

Por otro lado, las relaciones discursivas existentes en un texto, no siempre vienen evidenciadas por marcadores discursivos. Así resulta muy complicado definir dichas relaciones automáticamente. La postura de Marcu (1997) es realmente interesante pero, al intentar verificarla, nuestra experiencia empírica después del análisis del corpus deja ver que las relaciones discursivas no siempre vienen evidenciadas por marcadores.

Para observar la densidad de estos marcadores en los textos médicos de nuestro corpus, hemos realizado una serie de pruebas. En primer lugar hemos segmentado los textos en unidades discursivas mínimas (patrones núcleo-satélites) y hemos contado las relaciones establecidas por la RST (Mann & Thompson, 1988). En segundo lugar hemos llevado a cabo un conteo de los marcadores que unen o relacionan los elementos de estas relaciones discursivas. Para ello hemos tomado como modelo la clasificación de Portolés (1998), además de considerar algunas partículas gramaticales (preposiciones y conjunciones) que en ocasiones evidencian relaciones internas de las oraciones y que deben ser marcada al construir una estructura discursiva arbórea, como, por ejemplo, *desde, hasta, cuando, como*.... (Marcu, 1997b). En los ejemplos siguientes observamos un conector contraargumentativo, *por el contrario*, que relaciona dos elementos, y dos casos de elementos incrustados que dejan ver una relación de circunstancia temporal y modal, evidenciada por los marcadores *cuando* y *como,* respectivamente:

> Ej. 2. ["En el análisis provincial, Baleares, Palencia y Pontevedra presentaron prevalencias superiores al conjunto;][*por el contrario*, Tenerife, La Coruña y Murcia tuvieron niveles inferiores."]

Ej. 3. ["Los resultados de la RM-mielografía, [*cuando* generaban información nueva], se clasificaron también en relevantes y no relevantes."]

Ej. 4. ["La práctica de una política antibiótica restrictiva respecto a la utilización de antibióticos, [*como* se ha llevado a cabo en otros hospitales], reducirá significativamente la colonización y la CACD."]

De los diez textos de este corpus (18.098 palabras) el 1'64 % lo forman marcadores de relación discursiva (298 ocurrencias). Hemos realizado una tabla con las frecuencias de los marcadores en el corpus; a continuación se ofrece una muestra de los más representativos[3]:

marcador	*nº de ocurrencias*	*porcentaje*
y (e)	76	0'419 %
pero	32	0'176 %
si	29	0'160 %
cuando	19	0'104 %
según	14	0´077 %
aunque	13	0'071 %
como	12	0'066 %
por tanto	10	0'055 %

Tabla 1. Muestra de los marcadores discursivos más representativos en corpus médico.

De la segmentación de los textos en elementos discursivos, tomando para ello las relaciones de la RST, se han obtenido un total de 1059 relaciones. Si dividimos el número de relaciones discursivas entre el número de marcadores, llegamos a la conclusión de que hay un marcador para cada casi cuatro relaciones (3'55 relaciones). Llegados a este punto es necesario preguntarse qué ocurre con el resto de relaciones discursivas que no aparecen evidenciadas por medio de marcadores, ya que el análisis automático de las mismas resultaría inviable sin estas partículas.

5. CONCLUSIÓN: ESTRUCTURA INFORMATIVA Y SINTÁCTICA (3ER Y 4º NIVEL)

Lo que se quería demostrar en este artículo es que el marcaje de las relaciones discursivas (y de sus marcadores) juega un papel relevante a la hora de extraer los fragmentos más importantes de un texto, en este caso artículos científicos de medicina, para llegar finalmente a un resumen automático del mismo, pero que debemos tener en cuenta distintas facetas de los textos para poder llegar a una correcta representación de su totalidad. Creemos que aquí hay una carencia, y que para suplirla debemos acudir a un tercer y cuarto nivel: el de la estructura informativa y sintáctica. Este punto ya no lo trataremos en este trabajo, pero por aquí es por donde irán nuestras directrices futuras.

---

[3] Datos procedentes del Corpus Tècnic del IULA de la UPF (CT-IULA) obtenidos a través de BwanaNet en el período (02/2004).

Nuestro planteamiento parte de la idea de que la estructura informativa y sintáctica que subyace en los textos es básica a la hora de realizar el análisis de un documento para posteriormente llegar a un resumen del mismo. Así, tomaremos como base teórica la Teoría Comunicativa de Mel'cuk (2001), que consiste en la utilización de una jerarquía de oposiciones semántico-comunicativas, como son: *thematicity, givenness, focalization, perspective, emphasis, presupposedness, unitariness,* y *locutionality.*

Por otro lado, Mel`cuk (1998) utiliza una sintaxis de dependencias, que también aportará información a la hora de analizar el texto orientado al resumen. Así, la estructura informativa y la sintáctica están íntimamente relacionadas. Hay que tener en cuenta que "existe continuidad entre la sintaxis de la oración y la sintaxis –organización– del discurso" (Givon, 1979).

Por todo esto creemos que, a la hora de hacer un resumen, la información que no aportan el primer y segundo nivel de análisis (estructura textual y discursiva), pueden ofrecerla este tercer y cuarto nivel, el de la estructura informativa y sintáctica. Es aquí donde encontraremos los datos decisivos de análisis.

6. REFERENCIAS BIBLIOGRÁFICAS

ALONSO, Laura & Irene CASTELLÓN (2001): *Aproximació al Resum Automàtic per Marcadors Discursius, X-Tract WP 01/07*. Barcelona, Universitat de Barcelona.

BACH, Carme (2001): *Els connectors reformulatius catalans: anàlisi i proposta d'aplicació lexicogràfica*. Barcelona, IULA, Pompeu Fabra.

BARZILAY, Regina & Michael ELHAHAD (1997): "Using lexical chains for text summarization". Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.

BERGER, A. & V. MITTAL (2000): "A system for summarizing Web Pages". Proceedins of SIGIR'00. Atenas.

BHATIA, Vijay (1993): *Analyzing genre: Language Use in Professional Settings*. Londres: Longman.

BOGURAEV, Branimir & Christopher KENNEDY (1997): "Salience-based content characterization of text documents." Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.

BRANDOW, R., K. MITZE & L. RAU (1994): "Automatic condensation of electronic publications by sentence selection". Information Processing and Management, 31.

BRIZ, Antonio (2003): "Diccionario de partículas discursivas del español. Los resultados de un proyecto de investigación". Sociedad Española de Lingüística: XXXIII Simposio. Gerona.

BURGOS, R., J.A. CHICHARRO, & M. BOBENRIETH (1994): *Metodología de investigación y escritura científica en clínica*. Escuela andaluza de salud pública. Granada.

DUNNING, T. (1993): "Accurate Methods for the Statistics of Surprise and Coincidence. Computational Liguistics, 19.

EDMUNDSON, H.P. (1969): "New Methods in Automatic Extraction". En MANI & MAYBURY (1999), *Advances in Automated Text Summarization*. Cambridge.

GIVÓN, T. (1979): *Discourse and Syntax, Syntax and Semantics*. Nueva York: Academic Press.

GOLDSTEIN, J. , J. CARBONELL, M. KANTROWITZ, & V. MITTAL (1999): "Summarizing text documents: sentence selection and evaluation metrics". Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

HOBBS, Jerry R. (1978), "Why is discourse coherent?", nota técnica 176, SRI International Artificial Intelligence Center.

HOBBS, Jerry R. & Michael H. AGAR (1985): *The Coherence of incoherent discourse.* California: Stanford.

HOVY, Eduard (2003): "Information Retrieval, Question Answering, and Text Summarization". Seminario de Tecnologies de la llengua: recuperació de la informació. Univ. Internacional Menéndez Pelayo. Barcelona.

LIN, C. & HOVY E. (1997): "Identifying Topics by Position". Proceedings of the Applied Natural Language Processing Conference. Washington.

LÓPEZ ARROYO, Belén (2002): "La importancia de la estructuración externa de los *abstracts* en la enseñanza de los lenguajes con fines específicos". La enseñanza de lenguas en una Europa multicultural, Congreso Internacional de AESLA, Lugo.

LORÉS, Rosa (2002): "On the rhetorical structure(s) of abstracts". La enseñanza de lenguas en una Europa multicultural, Congreso Internacional de AESLA. Lugo.

LUNH, H. P. (1959): "The Automatic Creation of Literature Abstracts". En MANI & MAYBURY (1999), *Advances in Automated Text Summarization*. Cambridge.

MANN, William C. & Sandra A. THOMPSON (1988): "Rhetorical structure theory: Toward a functional theory of text organization". Text, nº 8, vol.3.

MARCU, Daniel (1996): "Building up rhetorical structure trees". Proceedings of the Thirteenth National Conference on Artificial Intelligence.

MARCU, Daniel (1997a): "From discourse structures to text summaries". Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.

MARCU, Daniel (1997b): "The Rhetorical Parsing of Unrestricted Natural Language Texts". Proceedings of the 35th Annual Meeting of the Association Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Madrid.

MARCU, Daniel (1998): *The rhetorical parsing, summarization, and generation of natural language texts*. Thesis. Department of Computer Science, University of Toronto.

MARCU, Daniel (2000): *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts, Institute of Technology.

MEL'CUK, Igor (1988): *Dependency Syntax*: *Theory and Practice*. Nueva York, Albany.

MEL'CUK, Igor (2001): *Communicative Organization in Natural Language. The semantic-communicative structure of sentences.* Amsterdam: John Benjamins.

ONO, K. , Kazuo SUMITA & Seiji MIIKE (1994): "Abstract generation based on rhetorical structure extraction". Proceedings of International Conference on Computational Linguistics.

PORTOLÉS, José (1998), *Marcadores del discurso*. Barcelona: Ariel.

PRADA, Juan José (2001): *Marcadores del discurso en español. Análisis y representación.* Uruguay.

SALAGER-MEYER , F. (1991): "Medical English Abstracts: How Well Are They Structure?". Journal of the American Society of Science. 42:7. 528-531.

TEUFEL, Simone & Marc MOENS (1997): "Sentence extraction as a classification task". Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.