Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Doctorado en Ciencias del Lenguaje y Lingüística Aplicada
Bienio 2002-2004

HACIA UN MODELO LINGÜÍSTICO DE RESUMEN AUTOMÁTICO DE ARTÍCULOS MÉDICOS EN ESPAÑOL

Proyecto de tesis

Iria da Cunha Fanego
Dirigido por Leo Wanner
Barcelona, septiembre 2005

ÍNDICE

ANEXOS

Anexo 1: artículo médico de nuestro corpus ("Visitas inapropiadas al servicio de urgencias de un hospital general"), resumen redactado por el autor del artículo y resúmenes automáticos fruto de la utilización de los seis sistemas comerciales seleccionados (*Extractor, Pertinence Summarizer, Summ-it, SweSum, GistSumm* y *Ms-Word Autosummarize*).

Anexo 2: tabla de contenidos seleccionados por el autor, los médicos y los lingüistas para el *Multidimensional Scaling*.

Anexo 3: Relaciones discursivas de la *Rhetorical Structure Theory***.**

Anexo 4: Árboles discursivos del artículo médico del Anexo 1 con relaciones de la *Rhetorical Structure Theory*.

Anexo 5: Árboles sintácticos del artículo médico del Anexo 1 con relaciones profundas de dependencia integradas en la *Meaning-Text Theory*.

Anexo 6: Ejemplos extraídos de nuestro corpus en los que funcionan las reglas sintáctico-discursivas.

Anexo 7: Apartado de Introducción de cuatro textos de nuestro corpus de contraste con sus respectivos resúmenes creados tanto por el autor como a partir de nuestros criterios.

ÍNDICE DE TABLAS Y GRÁFICOS

Resumen

En este proyecto de tesis se presenta una aproximación a un modelo de sistema de generación de resúmenes automáticos de artículos médicos en español basado en la información textual, léxica, discursiva y sintáctico-comunicativa que estos presentan. El objetivo de la tesis es el desarrollo de unas bases lingüísticas sólidas (basadas en las informaciones mencionadas) que sustenten dicho modelo, pero no llevaremos a cabo la implementación del mismo. Para el análisis de las relaciones discursivas empleamos la *Rhetorical Structure Theory* (Mann & Thompson, 1988) y para el análisis de las relaciones sintáctico-comunicativas utilizamos el marco de la *Meaning-Text Theory* (Mel'cuk, 1988).

1. Introducción

1.1. Motivación

En la actualidad, enormes cantidades de datos llegan a nuestras manos por diversos medios y con diferentes finalidades, sobre todo a partir de la extensión del uso de Internet. Sería imposible tratar de asimilar toda la información con la que nos encontramos, así que debemos seleccionar la que mejor se adapte a los intereses que nos ocupan y procesarla en el menor tiempo posible para tomar decisiones. Con esta finalidad nace el resumen automático, ya que la posibilidad de acceder a un resumen de cada documento que nos llegue nos ayudará a discernir si realmente merece la pena leerlo o no. En resumen, gracias a él podremos satisfacer este tipo de necesidades que surgen en la nueva era de la información y la tecnología. Teniendo en cuenta el avanzado estado de la lingüística computacional los sistemas de resumen automático tienen posibilidades reales de implementación. Así, poco a poco, la generación automática de resúmenes se ha convertido en un problema complejo sobre el que se está trabajando desde diversos puntos de vista. Las dos perspectivas principales son la estadística y la lingüística. Nosotros adoptamos la segunda de ellas, ya que defenderemos que este tipo de información es indispensable para llegar a un buen resumen automático. Aunque muchos de los sistemas actuales de generación de resúmenes automáticos explotan información lingüística de algún tipo en los textos, esto en ocasiones es insuficiente. Creemos que para poder alcanzar un resumen totalmente

coherente y que contenga los contenidos apropiados deben integrarse informaciones lingüísticas de varios tipos. Así, como iremos viendo a lo largo de este trabajo, buscaremos una perspectiva de integración de análisis textual, léxico, discursivo y sintáctico-comunicativo, para poder obtener una completa representación lingüística de los textos y posteriormente ofrecer un buen resumen de los mismos.

1.2. Ideas de partida

Como ya hemos comentado, los sistemas actuales de generación automática de resúmenes aún tienen carencias porque no explotan toda la información lingüística de los textos que se necesita para llegar a un resumen de alta calidad. Veremos más adelante que muchos se basan en técnicas estadísticas (y así el texto se convierte en una entidad sólo matemática) y muchos otros utilizan información lingüística de algún tipo concreto, sin considerar otras. Nosotros partiremos del análisis de los textos desde varias perspectivas: textual, léxica, discursiva y sintáctico-comunicativa. Por un lado, para explotar la información discursiva utilizaremos las relaciones de la *Rhetorical Structure Theory* (Mann & Thompson, 1988). Actualmente ya hay algunos sistemas que utilizan esta información lingüística para llegar al resumen automático (Marcu, 2000), lo cual nos servirá para comparar sus resultados con los nuestros. Por otro lado, para completar el análisis lingüístico previo a la realización del resumen automático, utilizaremos la perspectiva sintáctico-comunicativa, y para ello nos basaremos en la sintaxis profunda de dependencias, utilizando el marco de la *Meaning-Text Theory* (Mel'cuk, 1988). Actualmente se está iniciando en la Universidad de Montreal un proyecto sobre resumen automático en el que también se utiliza este tipo de información proveniente de la MTT (Bélanger & Kittredge, 2005) y con el que compartiremos algunas ideas.

Nuestro corpus está formado por textos del ámbito especializado de la medicina y en concreto del género *Artículo Original*. Esto se debe a que este tipo de textos posee una misma estructura fruto de las convenciones en el mundo de la medicina, mismos fenómenos informativos y comunicativos, y el objetivo compartido de la trasmisión del conocimiento científico. Además, estos textos están acompañados por el resumen del autor lo cual, como veremos más adelante en detalle, nos servirá de ayuda a la hora de validar nuestros resultados.

1.3. Trabajos previos

En una fase inicial de la investigación redactamos un artículo titulado "Importancia del marcaje de las relaciones discursivas para la generación automática de resúmenes" (da Cunha, 2004). En él se ofrecía una primera aproximación al resumen automático desde la perspectiva discursiva y se observaban algunas dificultades.

En 2005 llevamos a cabo el trabajo de línea (en el marco del doctorado en *Ciencias del Lenguaje y Lingüística Aplicada* del IULA-UPF) "Aproximación al resumen automático a partir de la integración de la perspectiva discursiva y sintáctica", en donde se ofrecía una propuesta de integración de análisis discursivo y sintáctico para solventar dificultades surgidas a partir de la simple utilización del discurso.

El artículo "Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria", ha sido aceptado para su presentación como *short paper* en el Workshop "Crossing Barriers in Text Summarization Research" que tiene lugar en el marco de la International Conference RANLP-2005 (*Recent Advances in Natural Language Processing*) en Borovets (Bulgaria). En él se ofrece una síntesis de nuestro modelo de resumen automático de artículos médicos en español basado en la integración de cuatro perspectivas lingüísticas: textual, léxica, discursiva y sintáctico-comunicativa.

Se ha enviado al séptimo Congreso de Lingüística General (Barcelona, 2006) un *abstract* para una propuesta de comunicación titulada "Resumen automático de artículos médicos en español: integración de técnicas de análisis textual, léxico, discursivo y sintáctico" que ha sido aceptada.

1.4. Interés del trabajo y posibles campos de aplicación

El interés específico de este trabajo, como ya hemos comentado, radica en la utilización de varias perspectivas lingüísticas (textual, léxica, discursiva y sintáctico-comunicativa) para llegar a un modelo lingüístico válido de generación de resúmenes automáticos de artículos médicos en español.

Otro de los intereses del estudio será validar la hipótesis de que la consideración de diferentes tipos de informaciones lingüísticas conlleva la mejora de la calidad de los resúmenes automáticos.

Una de las principales aplicaciones es que el modelo desarrollado pueda servir como base de implementación para crear un sistema que podrá ser de ayuda a médicos que se dediquen a la investigación en este ámbito y escriban artículos para publicar en revistas. Partimos de la base de que muchos de ellos necesitan economizar su tiempo y este sistema se lo permitiría, al ofrecerles un resumen automático de cada artículo que ellos escriban. Nuestro resumidor está orientado hacia especialistas en la materia ya que partimos de la premisa previa de que el resumen del autor es el ideal en este campo específico.

Otra de las finalidades de este sistema es su utilización por parte de alumnos de medicina de los últimos cursos o recién licenciados que quieran iniciarse en el mundo de la investigación. Estos jóvenes profesionales han dedicado la mayor parte de su tiempo al estudio de la medicina en sí, pero no tanto en aprender a redactar el tipo de resumen que les solicitarán las revistas de su área. Nuestro resumidor les servirá de ayuda en este aspecto ya que, cuando llegue el momento de escribir un artículo médico de investigación, podrán utilizarlo para resumir sus artículos y observar cómo se redacta un resumen adecuado. Se trataría de una herramienta didáctica para este colectivo, mediante la cual estos jóvenes aprenderían a seleccionar la información más importante de sus textos y a distribuirla de una manera ordenada de cara al resumen.

Finalmente, partimos de la idea de que el conjunto de reglas que desarrollaremos en la tesis será universal para ámbitos especializados, es decir, que se podrá aplicar con efectividad a otro tipo de campos diferentes de la medicina pero también específicos de alguna materia. Para corroborar esta hipótesis se realizarán experimentos en la tesis sobre textos de otros ámbitos. En concreto, se utilizarán los subcorpus especializados del Corpus Técnico del IULA (http://bwananet.iula.upf.edu/) cuyas materias son el derecho, la economía, el medioambiente y la informática.

2. Del resumen al resumen automático

2.1. Definición de resumen

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 Standards 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, 1993:78).

2.2. Parámetros de elaboración del resumen

Existen distintos parámetros (no excluyentes entre ellos) que deben ser tenidos en cuenta para elaborar un resumen: el documento a resumir (*input*), el resumen obtenido (*output*) y el propósito del mismo (*purpose*).

El *input* de un resumen 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 quiere obtenerse es una reestructuración coherente del texto o bien una extracción de los segmentos más relevantes del *input* (*abstract vs. extract*), o 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 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 de 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 que, por el contrario, tiene un conocimiento muy amplio, con lo cual el tipo de resumen también variará.

Estos parámetros aparecen reflejados esquemáticamente en la Tabla 1 (el subrayado indica las características del resumen que buscamos a partir de nuestro modelo):

INPUT	Único documento / Varios documentos
	Dominio específico / Ámbito general
	Texto monolingüe / Texto multilingüe
OUTPUT	*Abstract* / *Extract*
	Resumen neutral / Resumen evaluativo
PROPÓSITO DEL RESUMEN	Resumen indicativo / Resumen informativo
	Necesidades del autor / Necesidades del usuario
	Destinatario lego / Destinatario experto

Tabla 1. Parámetros de elaboración del resumen

Según observamos en la Tabla 1, el resumen al que queremos llegar es de un único documento, ya que la necesidad del lector en este campo es la de leer un resumen de un artículo concreto para observar sus objetivos y sus resultados específicos, no la de hacer un resumen que aúne varios artículos. El dominio con el que trabajamos es específico (medicina) y el texto monolingüe (español). En este estudio nos estamos refiriendo al *extract,* o lo que es lo mismo, al resumen creado con métodos de extracción ya que, después de la utilización de las técnicas que empleamos, nuestro resumen es una unión coherente de frases literales extraídas del texto original (aunque éstas puedan estar cortadas, lo que ocurrirá con frecuencia). De cara a la tesis examinaremos la posibilidad de utilizar técnicas de paráfrasis para la obtención de *abstracts*. El resumen que deseamos realizar es neutral, es decir, que se limita a la presentación de los datos médicos reflejados en el artículo original, sin una evaluación posterior. Además, el resumen es informativo (ya que necesitamos varios datos relevantes, no sólo hacernos una idea general del texto), se basa en las necesidades del autor (ya que no hay un usuario que haga consultas determinadas) y su destinatario es un experto en la materia (médicos, investigadores, estudiantes de medicina, etc).

2.3. El resumen automático: estado de la cuestión

Las investigaciones sobre resumen automático se están llevando a cabo desde diversas perspectivas y están adquiriendo una gran relevancia hoy en día. Desde sus inicios, en los años 60, se ha ido trabajando sobre ello utilizando diferentes técnicas y,

con el paso del tiempo, éstas han evolucionado, volviéndose más complejas. Muestras de esta evolución se observan en la página web actualizada periódicamente que mantiene Radev (www.summarization.com), con información sobre resumen automático (sistemas, congresos, recursos,...) y una extensa y completa bibliografía. Este autor es uno de los máximos exponentes de la investigación en este tema (Radev, 1999; Radev *et al*, 2001a; Radev *et al*, 2001b). Una de las obras que mejor refleja los avances en el campo es Mani & Maybury (1999), donde se lleva a cabo el primer compendio de los trabajos más importantes sobre resumen automático, organizándose en seis secciones: *Classical Approaches, Corpus-Based Approaches, Exploiting Discourse Structure, Knowledge-Rich Approaches, Evaluation Methods* y *New Summarization Problem Areas*. Posteriormente Mani (2001) prosigue con otra publicación en la que se ofrece una introducción en el tema, se explican definiciones básicas y se ofrece una amplia perspectiva de los métodos automáticos para generar *extracts* o *abstracts*, tanto de los que utilizan conocimiento lingüístico como estadístico.

Esta diferencia entre *extracts* o *abstracts* aparece reflejada en la definición de resumen automático que ofrece Sparck-Jones (2001): "a summary is a reductive transformation of a source text into a summary text by extraction or generation". Así pues, nos parece relevante, antes de explicar los diferentes tipos de técnicas de resumen automático, marcar esta distinción existente entre *extraction* (*extract*) y *generation* (*abstract*) que también señala Teufel & Moens (2002). Por un lado, un *extract* es un conjunto de pasajes (desde palabras sueltas a párrafos enteros) extraídos del texto *input* literalmente, que unidos forman un resumen. La mayoría de las aproximaciones al resumen automático toman esta concepción (Luhn, 1959; Edmunson 1969; Paice, 1990), ya que es muy útil en el campo de la recuperación de información para hacerse una idea del contenido de un texto (Mani *et al*, 1999). Por otro lado, un *abstract* es un texto generado nuevamente, producido por alguna representación interna que resulta después del análisis del *input*. Esta idea surge con el motivo de mejorar la calidad final del texto del resumen, y para ello se lleva a cabo un proceso posterior (Mani *et al*, 1999; Jing & Mckeown, 2000; Knight & Marcu, 2000).

2.3.1. Tipos de técnicas

Hovy (2003) divide los métodos actuales de generación automática de resúmenes, entendidos como *extracts*, en tres tipos: métodos superficiales, métodos de nivel medio y métodos profundos. A continuación comentaremos algunos de ellos.

2.3.1.1. Métodos superficiales

En cuanto a los métodos superficiales se incluyen pistas dadas por frecuencias de palabras (Luhn 1959; Edmunson, 1969), que se basan en la idea de que las entidades más importantes tienden a ser mencionadas más a menudo, con lo cual se atribuye más peso a las oraciones que contienen palabras frecuentes inusuales. Otro método superficial es el que se basa en la posición de determinados fragmentos, que se utiliza para géneros que tienen una estructura fijada, la cual se explota (títulos, secciones, *abstracts*,...). Por ejemplo, en un artículo periodístico las primeras líneas serán las más relevantes (Brandow *et al*, 1994; Lin & Hovy, 1997). Por otro lado, también pueden utilizarse como pistas los títulos, partiendo de la idea de que las palabras que estos contienen son importantes, con lo cual se añade puntuación a las oraciones que también las incluyen (Luhn, 1959). También es conocida la técnica que da importancia a las *cue phrases* (palabras clave), es decir, a las palabras o frases que indican centralidad como, por ejemplo, "es importante destacar que..." o "en conclusión..."(Edmundson, 1969). Por el contrario, existen algunas de estas frases que indican el poco peso para el resumen de la oración en la que se incluyen, como "por ejemplo..." (Teufel & Moens, 1999). Otros experimentos que explotan información lingüística utilizan el metadiscurso como posible indicación de relevancia: frases clave que se usan como convención en los artículos de investigación o determinados verbos que se utilizan con frecuencia para expresar conceptos importantes. Algunos de los autores que trabajan sobre resumen automático ya han explotado este tipo de marcas con buenos resultados (Pollock & Zamora, 1975; Paice, 1990). Varios de los métodos mencionados sirvieron para iniciar la investigación sobre el resumen automático.

2.3.1.2. Métodos de nivel medio

Los métodos de nivel medio incluyen técnicas de reconocimiento de cadenas léxicas, que enlazan elementos relacionados, encuentran cadenas y centralidad en oraciones y párrafos conectados, y estudian la correferencia (Barzilay & Elhahad, 1997; Boguraev & Kennedy, 1997). Cercano a estos planteamientos se encuentra el trabajo de Fuentes *et al.* (2004) en donde se presenta un sistema que permite resumir automáticamente documentos escritos en catalán (en concreto noticias de agencia) mediante extracción de fragmentos, explotando las propiedades de cohesión del texto a través de la detección de cadenas léxicas, de correferencia y de entidades nominales. Una aproximación previa a este trabajo se encuentra en Fuentes & Rodríguez (2002).

También hay métodos basados en la Máxima de Relevancia Marginal, los cuales seleccionan la oración central del texto y luego computan la relevancia marginal de las otras usando una fórmula MMR; posteriormente seleccionan la más relevante y la añaden al resumen, vuelven a computar la relevancia marginal del resto y de nuevo añaden la mejor al resumen; se detienen cuando la longitud del resumen es la deseada y reordenan las oraciones (Goldstein *et al*, 1999).

Muchas de las aportaciones en el campo del resumen automático no se centran en aspectos propiamente lingüísticos, sino estadísticos, por lo que los resultados aún dejan bastante que desear. 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 extracción de información (Berger & Mittal, 2000; Brandow *et al*, 1994; Dunning, 1993). En Kupiec *et al.* (1995) se desarrolla un sistema de resumen automático con técnicas de entrenamiento basado en la utilización de herramientas estadísticas que ofrece buenos resultados. El método concreto *Optimal Position Policy* emplea una técnica que "aprende" cuáles son las oraciones con más peso dado un corpus de entrenamiento (Luhn, 1959; Lin & Hovy, 1997). En Fuentes *et al.* (2003) se presenta un sistema de resumen que extrae la frase más importante de un texto mediante una aproximación de aprendizaje automático, comprimiéndola posteriormente mediante reglas manuales para obtener una frase gramaticalmente correcta.

2.3.1.3. Métodos profundos

Estos y otros ejemplos reflejan una perspectiva desde donde abordar el reto del resumen automático, pero hay otras aproximaciones que parten de la idea de que en este campo de investigación deben tratarse aspectos más lingüísticos. Así, podrían utilizarse métodos profundos como los que trabajan con técnicas basadas en la utilización de la estructura del discurso, que dan importancia a los componentes nucleares de las relaciones discursivas (Marcu, 1998; Ono *et al*, 1994). Este aspecto será uno sobre los que nos centraremos en nuestro trabajo ya que, como veremos, la representación discursiva de los textos será una de las informaciones (entre otras, como la sintáctico-comunicativa) que integraremos en nuestro modelo de sistema de resumen automático.

Observamos, pues, que los sistemas avanzados de generación automática de resúmenes de hoy en día 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 *Rhetorical Structure 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 basa 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 y/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. De hecho, en Bach (2005), por ejemplo, se señala que los marcadores de reformulación pueden ser indicadores de zonas de información especializada relevante. 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.

Existen otros trabajos que explotan aspectos lingüísticos de los textos. Entre otros, Teufel & Moens (2002) proponen un método para resumir artículos científicos (en concreto de lingüística computacional) basado en el estatus retórico de las afirmaciones que estos contienen. Presentan un algoritmo que utiliza una estructura retórica no jerárquica basada en siete categorías fijas: *aim, textual, own, background, contrast, basis, other*, para clasificar los contenidos de los artículos en cada una de ellas.

Las principales diferencias entre este método y el basado en las relaciones de la *Rhetorical Structure Theory* (Mann & Thompson, 1988) son que en el primero de ellos, por un lado, se parte de la idea de que la relevancia y la función de ciertas piezas pueden ser determinadas sin analizar la estructura jerárquica completa del texto y, por otro, que con su análisis buscan capturar el estatus retórico de una pieza del texto con respecto al mensaje completo.

Aproximaciones al resumen de noticias se observan en Alonso & Fuentes (2003b), en donde se presenta un resumidor que integra propiedades cohesivas del texto con relaciones de coherencia mediante la utilización de cadenas léxicas y de la estructura retórica y argumental obtenida mediante marcadores discursivos. Previamente en Alonso & Fuentes (2002) ya se había hecho una aproximación sobre cómo integrar métodos basados en cohesión con métodos basados en coherencia.

2.3.2. Resumidores para el español en Internet: sistemas disponibles y características

El objetivo de esta búsqueda es seleccionar de entre todos los sistemas de generación automática de resúmenes existentes en el mercado aquellos que se adapten a nuestros intereses. Necesitamos sistemas gratuitos disponibles en Internet que resuman textos en español para observar su funcionamiento, resultados y carencias. Llegado el momento de la evaluación de los resultados de nuestro modelo de resumidor (en la futura tesis) se realizarán comparaciones con estos sistemas mediante pruebas estadísticas, y por ello nos hemos limitado a seleccionar resumidores para el español. En la Tabla 2 observamos los más representativos.

Existen otros sistemas de generación de resúmenes automáticos, pero no están incluidos en esta tabla por varios motivos. El *Copernic* es un sistema bastante conocido para el inglés, francés, alemán y español, pero la versión de prueba, de la que puede disponer el usuario durante 30 días, sólo está disponible para las tres primeras lenguas. Otro sistema interesante sería el *InXight Summarizer Plus*, utilizado para varias lenguas (español, inglés, alemán, chino, etc.), pero que no tiene versión de demostración. Y finalmente, el sistema *Summarist* (con una versión demo on-line), creado por Hovy & Lin (*Information Sciences Institute* de la *University of Southern California*), que funciona para el español, inglés, francés, alemán e indonesio, no se encuentra activa actualmente (http://www.isi.edu/natural-language/projects/nlg-demonstrations.html).

NOMBRE y ACCESIBILIDAD	**IDIOMAS DISPONIBLES y URL**
Extractor (versión demo con posibilidad de instalación)	Español, inglés, francés, alemán, japonés, coreano. http://www.dbi-tech.com/dbi_extractor.asp
Pertinence Summarizer (versión demo on-line con contraseña de duración limitada)	Español, inglés, francés, alemán, italiano, portugués, japonés, chino, coreano, árabe, griego, noruego, ruso, holandés. http://www.pertinence.net
Summ-it (*System Quirk*) (versión demo on-line)	Cualquier lengua. http://www.computing.surrey.ac.uk/ai/SystemQ/
SweSum (versión demo on-line)	Español, inglés, francés, danés, alemán, sueco. http://swesum.nada.kth.se/index.html
GistSumm (versión demo con posibilidad de instalación)	Inglés, portugués, español. http://www.nilc.icmc.usp.br/~thiago/GistSumm.html
MS-Word Autosummarize (incluido en Word)	Cualquier lengua. Incluido en Word

Tabla 2. Sistemas de resumen automático en español disponibles en Internet.

En la Tabla 3 se ofrecen las principales características de los sistemas antes mencionados.

En el Anexo 1 se ofrece un artículo médico de nuestro corpus ("Visitas inapropiadas al servicio de urgencias de un hospital general"), el resumen redactado por su autor, y los resúmenes automáticos creados a partir de la utilización de los seis sistemas comerciales disponibles en Internet (*Extractor, Pertinence Summarizer, Summ-it, SweSum, GistSumm y Ms-Word Autosummarize*) con un parámetro máximo de longitud del 10%.

RESUMIDORES AUTOMÁTICOS	**Características del sistema**				
Nombre del sistema	Formato del texto	Longitud	Eliminación de *stop words*	Campos específicos	Métodos del resumen
Extractor	.txt .rtf .htm	Entre 3 y 30 frases (en demo 8 máximo)	sí	no	Algoritmos estadísticos y lingüísticos.
Pertinence Summarizer	.txt .html .pdf .rtf .doc	%	sí	sí	*Multilevel highlighting, keywords*, dominios específicos.
Summ-it	cualquiera	%	sí	no	Cohesión léxica.
SweSum	.txt .doc .html	% palabras caracteres	no	sí	Métodos estadísticos, lingüísticos y heurísticos.
GistSumm	.doc .txt	%	no	no	Estadística, *keywords*, TF-ISF, cohesión léxica.
Ms-Word Autosummarize	.doc	% oraciones palabras	no	no	Estadística.

Tabla 3. Características de los principales sistemas de generación automática de resúmenes en Internet

3. El género artículo médico

3.1. Estructura y contenido

Debido a nuestra intención de resumir artículos médicos en español creemos conveniente aclarar su estructura y contenidos habituales. Partimos de la base de que el resumen que acompaña al texto, redactado por el mismo autor, refleja correctamente los contenidos más relevantes del artículo original, lo cual nos será de ayuda a la hora de validar nuestros resultados.

3.1.1. Estructura IMRD

El ámbito de la medicina está cubierto por una gran variedad de géneros discursivos *(Artículos Originales, Notas Clínicas, Cartas al Director, Editoriales, Revisiones, Conferencias Clínicas, Conferencias Clínico-patológicas, Diagnóstico, Tratamiento*, etc.), pero en este estudio nos centraremos en el *Artículo Original*. Las revistas médicas solicitan a los autores de *Artículos Originales* que sigan un patrón determinado que debe mantener la siguiente estructura:

- Título
- Resumen en español
- Palabras clave en español
- Resumen en inglés (*abstract*)
- Palabras clave en inglés (*keywords*)
- Introducción
- Pacientes y métodos
- Resultados
- Discusión
- Agradecimientos (opcional)
- Bibliografía

Este tipo de textos sigue este patrón habitual solicitado por las revistas para su publicación. A continuación detallaremos los contenidos que deben incluirse en cada uno de los cuatro apartados principales que se caracterizan por seguir el orden lógico del

pensamiento científico: *Introducción*, *Pacientes y métodos*, *Resultados* y *Discusión*. Esta estructura recibe el nombre de "estructura IMRD". Para la especificación de los contenidos necesarios en cada una de las secciones seguimos las directrices de una de las revistas médicas de más prestigio y calidad en España: *Medicina Clínica.*

La *Introducción* de este tipo de artículos debe ser breve y proporcionar sólo la explicación necesaria para que el lector pueda comprender el texto que sigue a continuación. No debe contener tablas ni figuras. Debe incluir de forma clara el/los objetivo/s del trabajo. Siempre que se pretenda publicar una observación muy infrecuente debe precisarse en el texto el método de búsqueda bibliográfica, las palabras clave empleadas, los años de cobertura y la fecha de actualización.

En el apartado de *Pacientes y métodos* se indican el centro donde se ha realizado el experimento o la investigación, el período de duración, las características de la serie estudiada, el criterio de selección empleado y las técnicas utilizadas, proporcionando los detalles suficientes para que una experiencia determinada pueda repetirse sobre la base de esta información. Deben describirse con detalle los métodos estadísticos.

En la sección de *Resultados* deben relatarse, que no interpretar, las observaciones efectuadas con el método empleado. Estos datos deben exponerse en el texto con el complemento de las tablas y figuras.

Finalmente, en el cuarto y último apartado de la estructura IMRD, la *Discusión*, los autores tienen que exponer sus propias opiniones sobre el tema. Destacan aquí: 1) el significado y la aplicación práctica de los resultados; 2) las consideraciones sobre una posible inconsistencia de la metodología y las razones por las cuales pueden ser válidos los resultados; 3) la relación con publicaciones similares y comparación entre las áreas de acuerdo y desacuerdo, y 4) las indicaciones y directrices para futuras investigaciones. No deben efectuarse conclusiones. Además debe evitarse que la discusión se convierta en una revisión del tema y que se repitan los conceptos que hayan aparecido en la *Introducción*. Tampoco deben repetirse los resultados del trabajo.

3.1.2. Subtítulos del artículo médico

Partimos de la idea de que, a la hora de escribir un resumen de un artículo médico de este tipo, debe seleccionarse cierta información de cada uno de sus apartados (*Introducción, Pacientes y métodos*, *Resultados* y *Discusión*) para que éste siga el

proceso lógico del pensamiento científico. Esta idea de explotación de la estructura textual en resumen automático también se sigue en el trabajo de Bélanger (2005).

Para que nuestro sistema reconozca la existencia de estos apartados nos basaremos en el reconocimiento de sus títulos y, para ello, hemos realizado un estudio sobre los títulos de los apartados de nuestro corpus de 20 artículos médicos, llegando a la conclusión de que, aunque estos están en principio fijados, hay ligeras variaciones entre ellos, que deben ser tenidas en cuenta para que el sistema pueda reconocerlos con efectividad.

En la Tabla 4 observamos los diferentes subtítulos encontrados en nuestro corpus de 20 artículos médicos con sus respectivas frecuencias (tanto en los resúmenes como en los artículos). Debemos tener en cuenta que los textos que pretendemos resumir, como veremos más adelante, estarán lematizados, con lo cual no habrá ningún problema a la hora de distinguir entre singular y plural. Ej. *Pacientes y métodos* vs. *Paciente y método*.

Apartado	**Resumen**	**Frec**	**Artículo**	**Frec**
1	*Fundamento* *Objetivo*	18 2	---	20
2	*Pacientes y métodos* *Sujetos y método* *Material y método* *Métodos* *Población y métodos*	8 4 2 3 3	*Pacientes y métodos* *Sujetos y método* *Material y método* *Método* *Población y método*	11 3 2 1 3
3	*Resultados*	20	*Resultados*	20
4	*Conclusiones*	20	*Discusión*	20

Tabla 4. Subtítulos del artículo médico: formas y frecuencias

En primer lugar se observa que el apartado de *Fundamento* del resumen se denomina siempre así, aunque hay dos casos en los que los autores se decantan por *Objetivo*. Debemos tener en cuenta que, en el artículo en sí, el título de este apartado no se ve reflejado explícitamente porque se sobreentiende.

En segundo lugar se observa que el apartado de *Pacientes y métodos* presenta una gran variación en cuanto a su denominación se refiere. Por un lado, en el resumen predomina la forma *Pacientes y métodos*, con ocho ocurrencias, ya que es la manera más habitual de referirse a este apartado. También nos encontramos con otras formas menos frecuentes, aunque no por ello menos importantes, ya que debemos encontrar

todas las posibles variantes para que el resumidor las reconozca. Sería el caso de *Sujetos y método* (con 4 ocurrencias), *Material y método* (con 2 ocurrencias), *Métodos* (con 3 ocurrencias) y *Población y métodos* (con 3 ocurrencias). Por otro lado, en el artículo también predominan las formas *Pacientes y métodos*, con 11 ocurrencias. Además encontramos las formas *Sujetos y método* (con 3 ocurrencias), *Material y método* (con 2 ocurrencias), *Método* (con 1 ocurrencia) y *Población y método* (con 3 ocurrencias). Son ligeras diferencias para expresar el mismo concepto, pero todas ellas deben ser tenidas en cuenta para un efectivo reconocimiento de los títulos.

En tercer lugar, se encuentra el apartado de *Resultados*, que no presenta ninguna variación en cuanto a su denominación, ni en el resumen ni en el artículo.

Por último, el cuarto apartado se denomina en los 20 resúmenes *Conclusión*, mientras que en el artículo encontramos la forma *Discusión*. Esta diferencia de denominación en este apartado se debe a que también varían los datos que se incluyen en él, dependiendo de si se trata del artículo o del resumen. En la *Discusión* no se explicitan exactamente las conclusiones del trabajo desarrollado, mientras que en el apartado de *Conclusión* del resumen sí deberían reflejarse. Esto deberá ser tenido en cuenta en la tesis de cara al desarrollo de nuestro resumidor automático.

3.1.3. Unidades léxicas representativas del artículo médico

Partimos de la base de que al intentar resumir textos de un ámbito especializado y restringido hay unidades léxicas importantes que aparecen con cierta frecuencia en dichos textos, ofreciendo indicaciones acerca de cuáles son los contenidos importantes. En cada ámbito especializado habrá ciertas unidades que cumplan esta función. Así, por ejemplo, en el ámbito del derecho consideraríamos unidades representativas *prohibir, infringir, cometer, fundar, ejercer*, etc. En el tipo de artículos médicos que nosotros queremos resumir también hay determinadas unidades léxicas que indican que los fragmentos que las contienen son relevantes a la hora de redactar un resumen. En el apartado 8.1.2. observaremos más datos en este sentido, pero se trataría básicamente de unidades nominales que intuitivamente reflejan conceptos importantes para un resumen de este tipo (como *objetivo, objeto, propósito, intención, resumen, conclusión, resultado*) o unidades verbales, con frecuencia alta y no auxiliares o demasiado comunes (como *asociar, analizar, presentar, evaluar, relacionar, aportar,*

estudiar, valorar). En el apartado 8.1.2. se encuentra la lista provisional de unidades léxicas seleccionadas.

Mediante su detección, el sistema de resumen dispondrá de un primer *output* de oraciones (las que incluyan las unidades léxicas que nosotros determinemos) que posteriormente se contrastará y refinará mediante otras técnicas lingüísticas que se explicarán más adelante.

3.2. El resumen científico

Una vez hemos concretado los contenidos que deben incluirse en cada uno de los apartados de este tipo de artículos médicos y hemos explicado la importancia de sus títulos y de ciertas unidades léxicas relevantes, nos centraremos ahora en el resumen en sí.

Como ya hemos comentado en el apartado 2.1, el *American National Standards 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: 78).

Según las directrices de *Medicina Clínica* el resumen que acompaña al artículo debe adjuntarse en español y en inglés, y su contenido también debe dividirse en cuatro apartados y mantener la estructura IMRD del texto original: *Introducción*, *Pacientes y métodos*, *Resultados* y *Discusión*, para expresar de forma más breve pero con igual eficacia el proceso lógico del pensamiento científico. En cada uno de ellos deben describirse, respectivamente, el problema motivo de la investigación, la manera de llevarla a cabo, los resultados más destacados y las conclusiones que derivan de los resultados. La extensión del resumen en los *Artículos Originales* debe ser como máximo de 250 palabras y en los *Originales Breves* de 180 palabras aproximadamente.

3.3. Resumen del autor como punto de referencia

Consideramos que el resumen redactado por el mismo autor del artículo es el adecuado para este tipo de textos. Para justificarlo esbozamos una serie argumentos teóricos y, posteriormente, realizamos una prueba estadística que en principio los validará.

3.3.1. Argumentación teórica sobre el resumen del autor como punto de referencia

Creemos que el resumen que un autor redacta a partir de un artículo que ha escrito previamente es válido por varios motivos.

El primero de ellos es que el resumen del autor está orientado a especialistas en la materia, y al ser él mismo uno de ellos, sabrá con exactitud cuáles son los contenidos que debe incluir en el resumen, es decir, los más relevantes, y de los que debe prescindir por tratarse de información innecesaria. Estas consideraciones enlazan con Cabré (2002:89):

"Los textos devienen así producto de operaciones lingüístico-cognitivas realizadas en unas determinadas circunstancias discursivas. Estas circunstancias implican el emisor y el receptor (tipo de emisor y receptor, intenciones y nivel de conocimiento que ambos poseen sobre el tema), la situación (el medio en que se produce la comunicación y el sistema de transferencia utilizado), el propósito y las expectativas del emisor y receptor con relación a su interacción. Cada uno de estos factores posee un determinado valor en cada acto comunicativo y su conjunto explica la configuración de un tipo de texto que pretende ser adecuado a las circunstancias en las que se produce."

En segundo lugar, es el mismo autor quien escribe tanto el artículo como el resumen, con lo cual estos estarán estrechamente ligados y se corresponderán en cuanto a estructura y contenido.

En tercer lugar, la propia revista en donde se publicará el artículo (aquí *Medicina Clínica*) facilita a los autores una guía de estilo para publicaciones médicas (http://www.doyma.es/revista/info/pdf/002Normas.pdf).

Como cuarto motivo, ofrecemos el ya comentado en apartados anteriores, es decir, la obligación de los autores de seguir una estructura consensuada basada en cuatro apartados fijados (estructura IMRD). Esta idea enlaza con van Dijk (1989:165):

"[...] la aceptabilidad de la publicación depende de una serie de criterios que exigen métodos e informes adecuados"

Finalmente, debemos ser conscientes de que el artículo se publicará *con* el resumen, es decir, que el redactado de ambos debe ir al unísono para que tenga

coherencia. Caso distinto sería el de los resúmenes que se envían a congresos, ya que posteriormente el artículo podría variar y no seguir exactamente las ideas del resumen inicial.

3.3.2. Validación empírica

Para corroborar la hipótesis teórica de que el resumen del autor es adecuado, hemos realizado un experimento en el cual han colaborado tres médicos y tres lingüistas. El experimento ha consistido en ofrecer por separado a estas seis personas cinco artículos médicos (sin los resúmenes) y pedirles que seleccionasen en cada texto los contenidos que considerasen indispensables para construir un buen resumen del mismo (ofreciéndoles unos parámetros de longitud máxima). Partimos de la idea de que los médicos tienen la base teórica para seleccionar los contenidos más adecuados, y de que los lingüistas tienen el conocimiento "lingüístico" para seguir la estructura del texto y deducir qué informaciones escoger. Como ya hemos mencionado, consideramos que el resumen del autor es adecuado, y la semejaza de los resúmenes de los autores con los de los médicos constataría esta afirmación.

Así, hemos comparado los contenidos del resumen del autor con los contenidos seleccionados por estos tres médicos y tres lingüistas, y se ha constatado que, efectivamente, el autor coincide mayoritariamente con los tres primeros. Para cuantificar los resultados hemos utilizado el programa estadístico *Statgraphics*, mediante el cual se ha realizado un *Multidimensional Scaling* para comprobar el grado de coincidencia entre los resúmenes del autor, de los médicos y de los lingüistas. Los resultados del experimento se muestran en el Gráfico 1. Se observa que los tres médicos y los autores de los artículos seleccionan contenidos muy similares a la hora de redactar los resúmenes de los artículos, mientras que los lingüistas no se asemejan ni siquiera entre ellos. Mediante esta validación estadística tenemos una primera prueba empírica de que el resumen del autor es correcto, ya que otros tres especialistas en medicina así lo constatan. La tabla de contenidos seleccionados por el autor, los médicos y los lingüistas puede observarse en el Anexo 2.

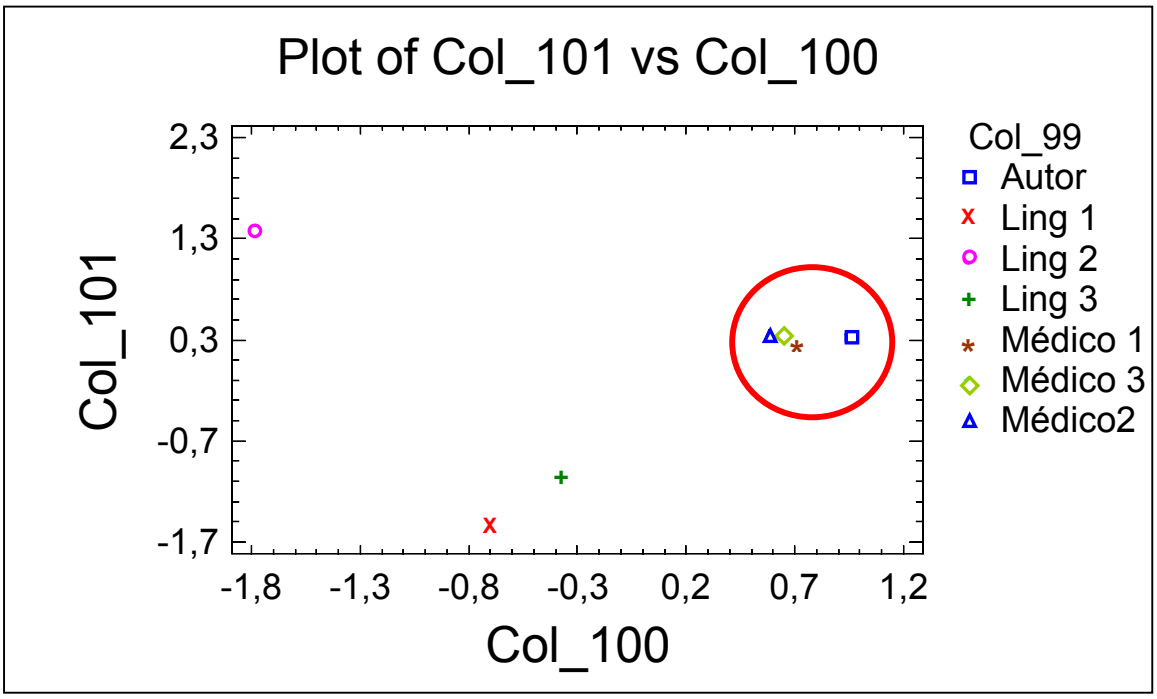

Gráfico 1. *Multidimensional Scaling* a partir de los contenidos escogidos para los resúmenes

Las instrucciones que se ofrecieron a los seis informantes fueron las siguientes:

• **Subraye** en cada texto los fragmentos que considere *indispensables* para construir un buen resumen del mismo (excepto títulos y subtítulos):

- Originales = máximo 20 líneas subrayadas (aprox.)
- Originales Breves = máximo 15 líneas subrayadas (aprox.)

Después de realizar este experimento y de constatar la coincidencia entre los resúmenes de los médicos y la gran diferencia entre los fragmentos seleccionados por estos dos colectivos de profesionales (médicos y lingüistas), se ha observado qué tipo de contenidos ha escogido cada uno de ellos. Notamos una tendencia generalizada en los lingüistas a incluir en su resumen demasiada información del apartado de *Introducción*. Creemos que esto es debido a su falta de formación en medicina, la cual les lleva a seleccionar definiciones, datos históricos, trabajos anteriores al del artículo que se quiere resumir, etc. En cambio, los médicos no incluyen este tipo de informaciones debido a que éstas se sobreentienden al poseer una formación específica en este ámbito. Otro dato que nos llama la atención es que los médicos seleccionan información

importante de cada uno de los apartados, y no así los lingüistas, que en algunos casos descuidan información de alguno de ellos. Observamos además que los tres médicos suelen seleccionar información numérica, sobre todo en el apartado de *Pacientes y métodos* y *Resultados*, mientras que los lingüistas tienden a seleccionar información no numérica, con más explicaciones y menos cifras. Finalmente, vemos que los lingüistas seleccionan bastantes contenidos del apartado de *Conclusión*, mientras que los médicos ofrecen normalmente una conclusión mucho más breve.

Todas estas observaciones nos hacen ver que la formación concreta del colectivo que vaya a resumir un texto influye a la hora de hacerlo. Es decir, que aunque podamos creer de antemano que la formación de un lingüista pueda ser suficiente a la hora de seleccionar la información importante (ya que se supone que éste posee capacidades para la correcta visualización de la estructura del discurso, para encontrar conectores discursivos que guíen dicha estructura, para entender y asimilar la estructura textual e, incluso, sintáctica) finalmente esto no es así. Puede ser que los lingüistas seleccionen información relevante, pero no es la necesaria para este tipo concreto de resumen, teniendo en cuenta el receptor. Son los resúmenes de los tres médicos los que se asemejan entre ellos y se acercan más al del autor, considerado éste como el resumen adecuado en este ámbito. Como conclusión del análisis, llegamos a la constatación de que para la realización de esta tarea se necesita ser un profesional en el ámbito al cual pertenecen los textos que se desee resumir. Los médicos, mediante su conocimiento y experiencia, son los más indicados para llevar a cabo un resumen correcto sobre un texto de medicina. Esto nos hace pensar, como vía de trabajo, en que no sólo se necesita información lingüística de un tipo (como la utilizada por Marcu mediante la RST) a la hora de resumir este tipo de textos, ya que no ofrecería buenos resultados. Necesitamos un análisis más profundo, además de la utilización de unidades léxicas representativas del ámbito temático en el que estamos trabajando.

De todas maneras debemos tener en cuenta de que habrá ocasiones en las que el resumen del autor pueda no ser el más indicado, ya que hay autores que no tienen suficiente capacidad para condensar las ideas más relevantes de sus propios artículos, y por tanto, el resumen podrá ser de una calidad inferior a la esperada. Esta situación no es la habitual, pero debemos tenerla en cuenta para futuros experimentos.

4. Marco teórico

Para poder entender las hipótesis que se plantearán más adelante creemos conveniente situarnos de manera general en el marco teórico de donde partimos. Por tanto, haremos una breve presentación de las ideas principales de cada uno de los dos marcos, con la idea de que, una vez asimilados los postulados principales de ambos, pueda entenderse la perspectiva de integración que aquí se propone.

4.1. La estructura discursiva

En este apartado explicaremos, en primer lugar, las bases de la *Rhetorical Structure Theory* (Mann & Thompson, 1988) como teoría discursiva tradicional y, en segundo lugar, daremos cuenta de la aplicación que Marcu (2000) ha hecho de cara al resumen automático.

4.1.1. Rhetorical Structure Theory tradicional (Mann & Thompson)

Para el análisis de los textos desde la perspectiva discursiva, tomaremos como base teórica la *Rhetorical Structure Theory* (RST) de Mann & Thompson (1988). La RST se creó en el marco de estudios de generación automática de textos. Un grupo de investigadores del *Information Sciences Institute* en la Universidad del Sur de California, percibieron que no había ninguna teoría de la estructura del discurso que ofreciera suficientes detalles para programar un generador automático de textos. Para solventar esta necesidad se creó la RST, que tiene validez en la actualidad como 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 discursivos 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, lo que nos lleva a enlazar esta teoría con el ámbito de la generación de resúmenes automáticos.

El esquema estructural más frecuente es el de dos unidades de texto (casi siempre adyacentes, aunque hay excepciones) relacionadas de tal manera que una de ellas tenga un papel específico con respecto a la otra. Un ejemplo es el de una afirmación que aporta una información básica acerca de algo seguida de una información adicional sobre lo mismo (ver Gráfico 2). La RST establece en este caso una relación de *Elaboración* entre las dos unidades. La relación también establece que la afirmación es más importante en el texto que la información adicional, de tal manera que la afirmación se convierte en el núcleo de la relación y la información adicional en su satélite. No hay reglas absolutas con respecto al orden de las unidades núcleo y satélite, aunque en la mayor parte de las relaciones puede encontrarse un orden preferido. Existen otras relaciones similares: *Evidencia*, *Fondo*, *Preparación* o *Concesión*. El listado de las relaciones discursivas establecidas por Mann & Thompson (1988) se encuentra en el Anexo 3.

En el caso de relaciones que no presentan una unidad central con respecto a los propósitos del autor, la relación se denomina *Multinuclear*. Un ejemplo lo constituye la relación *Multinuclear* de *Lista*, en donde se realiza una enumeración de varios elementos que tienen la misma importancia. Otras relaciones de este tipo son *Contraste*, *Secuencia* o *Unión* (ver Anexo 3).

En el Gráfico 2 observamos un ejemplo de un fragmento de estructura arbórea con relaciones de la RST, en donde se ofrece una relación de *Concesión*, otra de *Elaboración* y una última relación *Multinuclear* de *Lista*: "Es posible que algunas visitas consideradas adecuadas por el PAUH pudieran haber sido resueltas en atención primaria, pero los médicos del servicio de urgencias las trataron como si fuesen apropiadas: solicitaron exploraciones complementarias de los pacientes o les administraron tratamientos parenterales."

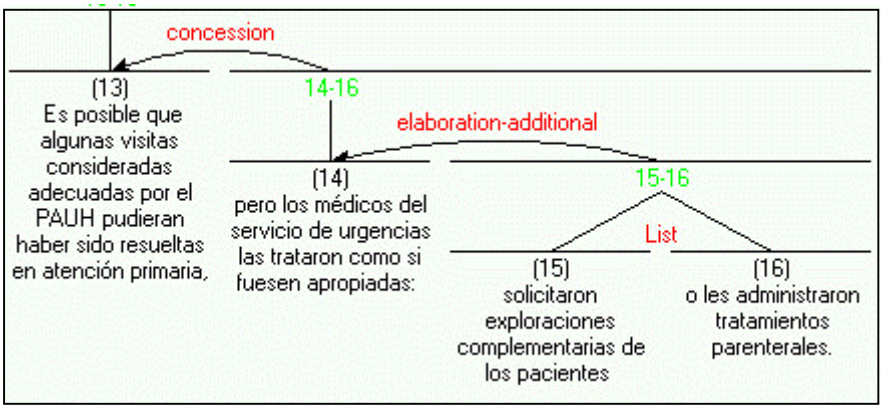

Gráfico 2. Fragmento de estructura arbórea con relaciones de la RST

En el presente trabajo partimos de la noción de Unidad Discursiva Mínima (UDM) que toman Carlson & Marcu (2001) para anotar su corpus discursivamente con relaciones de la Rhetorical Structure Theory[1]. Así (Carlson & Marcu, 2001:3):

"Our goal was to find a balance between granularity of tagging and ability to identify units consistently on a large scale. In the end, we chose the clause as the elementary unit of discourse, using lexical and syntactic clues to help determine boundaries. A few refinements to this basic principle are enumerated below, with reference to the section of the manual that discusses the phenomenon in more detail."

Los refinamientos a los que se refieren están indicados en su manual, pero se refieren a casos del tipo:

- Oraciones subordinadas que son sujetos u objetos del verbo principal no son tratadas como UDM.
- Oraciones subordinadas que son complementos del verbo principal no son tratadas como UDM.
- Las oraciones coordinadas se dividen en UDM diferenciadas, mientras que las frases verbales coordinadas no lo hacen.

Hay otras restricciones que, como hemos comentado, están reflejadas en Carlson & Marcu (2001) y que nosotros también adoptamos. Concretaremos ahora una de ellas en la que discrepamos:

- Los complementos de verbos de atribución (actos de habla u otros actos cognitivos) son tratados como UDM.

En nuestro trabajo estos complementos no serán tratados como tal, sino que se unirán al verbo ya que, como veremos más adelante, cuando realicemos el análisis sintáctico de los textos, esta unidad será el actante II. Veamos un ejemplo de Carlson & Marcu (2001):

[1] Este corpus está formado por 385 documentos de *American English* seleccionados del *Penn Treebank* (Marcus *et al*, 1993); las bases para anotarlo están ya esbozadas en Marcu (1999).

[Bush indicated] [there might be "room for flexibility" in a bill] [to allow federal funding of abortions for poor women] [who are victims of rape and incest.]

En nuestro trabajo esté fragmento se analizaría de la siguiente manera:

[Bush indicated there might be "room for flexibility" in a bill] [to allow federal funding of abortions for poor women] [who are victims of rape and incest.]

4.1.2. Aplicación de la RST cara al resumen (Marcu)

Marcu (1996), basándose en la RST, 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, con una orientación hacia la generación automática de resúmenes. Además lleva a cabo un prototipo de analizador discursivo automático, basado en gran medida en el uso de marcadores discursivos. Una vez creada automáticamente la estructura discursiva de un texto en términos de la RST se aplicará un algoritmo que proporcione un peso y un orden a cada elemento discursivo de la estructura (cuanto más alto esté el elemento en la estructura, más peso tendrá, y a la inversa), seleccionando para el resumen los elementos con mayor peso, y eliminando aquellos que tengan el peso más bajo. Dependiendo de la longitud que se desee para el resumen se escogerán más o menos elementos, pero siempre siguiendo el orden fijado por el algoritmo. En palabras de Marcu (2000:192):

"[...] a discourse-based summarization program that takes two arguments: a text and a number *p* between 1 and 100. The program first uses the cue-phrase-based rhetorical parser in order to determine the discourse structure of the text given as input. It then applies formula 9.1 and determines a partial ordering on the elementary and parenthetical units of the text. It then uses the partial ordering in order to select the *p%* most important textual units of the text."

Marcu (1998) aumenta el listado inicial de relaciones discursivas de la RST, ya que en realidad éste puede ampliarse o reducirse en función de las necesidades del usuario. En este trabajo nos hemos limitado a manejar las 26 relaciones discursivas

expuestas en un principio por Mann & Thompson (1988). Así, estas relaciones son las que analizaremos en los textos, teniendo en cuenta que, de momento, la diferencia existente entre *Resultado voluntario* e *involuntario* y *Causa voluntaria* e *involuntaria* no es pertinente (ver Anexo 3), ya que cara nuestros propósitos con respecto al resumen automático no parece que esta distinción sea motivo de elección o descarte de los fragmentos que las contengan.

4.2. La estructura sintáctica

Desde hace años se ha venido considerando que existen dos formas de análisis sintáctico: el análisis de constituyentes y el análisis de dependencias. En este estudio se utilizará el segundo de ellos; en concreto emplearemos la sintaxis profunda de dependencias integrada en la *Meaning-Text Theory* (MTT) de Mel'cuk (1988) ya que, como veremos, es una manera sencilla y efectiva de crear la estructura sintáctica de los textos y, además, nos permitirá algo importante de cara a nuestro trabajo, como es encontrar las relaciones de dependencia entre unidades léxicas. Comenzaremos este apartado observando una panorámica general de este tipo de sintaxis y explicando el porqué de su selección para este proyecto y no de la más tradicional de constituyentes; a continuación explicaremos en qué consiste esta sintaxis a partir del modelo de la MTT.

4.2.1. Las relaciones de dependencias

El término de *teoría de dependencias* se encuentra frecuentemente en la bibliografía y fue acuñado por Hays (1960, 1964); además también suele emplearse el término *gramática de dependencias* (Robinson, 1970). Las dependencias, como un modo formal de representar la estructura sintáctica de las oraciones, han sido utilizadas durante décadas por estudiosos de la sintaxis, y su apogeo llegó con el trabajo de Tesnière (1959). Posteriormente la sintaxis de constituyentes, acuñada con anterioridad por Bloomfield (1933), fue ganando terreno y desplazó a la de dependencias, relegándola a un segundo plano, sobre todo por ser la única forma de representación sintáctica empleada en los trabajos de Chomsky (1965) en su Escuela Generativa-Transformacional.

Posteriormente, Mel'cuk (1988) intenta sugerir un lenguaje formal artificial para describir oraciones naturales a nivel sintáctico y, concretamente, reivindica que el

formalismo de dependencias se adapta mucho mejor a la descripción de la estructura sintáctica que el de constituyentes. Hay varios puntos principales en los que la sintaxis de dependencias difiere de la de constituyentes (Mel'cuk, 1988), pero nosotros sólo haremos hincapié en aquellos que justifiquen nuestra elección de la sintaxis de dependencias frente a la de constituyentes de cara al resumen automático.

El primero de ellos es la utilidad de relaciones frente a constituyentes. Por un lado, en el llamado sistema o método de constituyentes la principal operación lógica es la inclusión de elementos en conjuntos, así estos pertenecen a una oración o a una categoría. Según esta aproximación, una oración es segmentada en constituyentes, cada uno de los cuales es consecuentemente segmentado. Así, esto favorece un punto de vista analítico. Por otro lado, la aproximación de dependencias se centra en las relaciones entre las unidades sintácticas últimas, es decir, entre las palabras. La principal operación aquí consiste en establecer relaciones binarias. Según esta idea, una oración se construye a partir de palabras unidas por dependencias. Así, esto favorece un punto de vista sintético, mucho más acorde con nuestros intereses sobre resumen automático, ya que nosotros necesitamos la composición de un todo por la unión de sus partes, y no una división y separación de las partes que forman un todo para llegar a conocer sus elementos. Un árbol de constituyentes puro no permite representaciones naturales de relaciones asimétricas de dependencia y no permite reflejar una distinción tan importante como la que hay entre un *government* (palabra principal) y su *dependent* (palabra que depende del núcleo), tan importante de cara al resumen. Para nuestros propósitos es muy importante tener en cuenta estas relaciones asimétricas, ya que en este trabajo buscamos integrar la sintaxis con el discurso, donde éstas también son de gran importancia, y una sintaxis de constituyentes no nos ofrecería esta posibilidad de integración. Nosotros queremos marcar las relaciones existentes entre elementos y el tipo de elemento que forma parte de esa relación, ya que lo que haremos posteriormente es solapar la estructura discursiva con la sintáctica. Cuando en el análisis de los textos coincidan determinados tipos de relaciones habrá elementos de esas relaciones que se eliminarán para el resumen. Por ejemplo, como veremos más adelante, cuando en una determinada secuencia haya un satélite de una relación de *Elaboración* en el análisis discursivo, y en esa misma secuencia, pero al realizar el análisis sintáctico, haya un adjunto apenditivo (APPEND), habrá que eliminarla por no aportar información relevante. Estos fenómenos no podrían ser observados si utilizásemos para el análisis

sintáctico una sintaxis de constituyentes ya que en ésta los elementos no se relacionan, sino que están unos dentro de otros, como en cajas.

En segundo lugar, en un árbol de constituyentes muchos nodos son no-terminales, es decir, que representan agrupaciones sintácticas o frases. Un árbol de dependencias, por el contrario, contiene sólo nodos terminales, y en él las palabras estructurales no se representan, pudiendo aparecer ciertos lexemas ficticios, funciones léxicas, símbolos idiomáticos, etc., aunque todos estos no necesariamente tengan un equivalente directo en la oración. Además, cualquier oración puede ser fácilmente especificada en un árbol de dependencias si se necesita, indicando su *government* y completando el subárbol que cuelga de él. Este procedimiento nos parece el más coherente, ya que si tenemos en mente llevar a cabo un resumen, deberemos ser lo más explícitos posible, analizando hasta la última unidad léxica por sí sola, porque si no aporta información esencial ésta deberá ser eliminada.

Finalmente, la tercera razón por la que en este estudio no se utilizará una sintaxis de constituyentes es que ésta no puede especificar el tipo de unión sintáctica existente entre dos elementos mientras que, por otro lado, una sintaxis de dependencias pone especial énfasis en especificar con detalle cualquier tipo de relación entre dos elementos. Esto es de especial relevancia para nuestros propósitos ya que, en muchas ocasiones, descartaremos o conservaremos determinados fragmentos de texto para el resumen dependiendo del tipo de unión sintáctica que exista.

4.2.2. Meaning-Text Theory: integración de la sintaxis de dependencias

Una vez aclarado esto, ya podemos explicar en qué se basa la sintaxis de dependencias, que se integra en la *Meaning-Text Theory* de Mel'cuk (1988).

Esta teoría presupone que un Modelo Sentido-Texto es un dispositivo lógico o conjunto de reglas que tiene como estructura general:

semántica		**sintaxis**		**morfología**		**fonología**
$\{Rsem_i\}$	**<===>**	**$\{RSint_k\}$**	**<===>**	**$\{RMorf_l\}$**	**<===>**	**$\{RFonet_j\}$**

y como estructura detallada:

Semántica **Sintaxis profunda** **Sintaxis superficial**

{Rsem$_i$} **<===>** **{RSintP$_{k1}$}** **<===>** **{RSintS$_{k2}$}** **<===>**

Morfología profunda **Morfología superficial** **Fonología**

{RMorfP$_{l1}$} **<===>** **{RMorfS$_{l2}$}** **<===>** **{RFonet$_j$}**

La sintaxis de dependencias evidenciada mediante estructuras arbóreas es el procedimiento que Mel'cuk utiliza para reflejar la sintaxis de la lengua dentro del modelo de la MTT. En concreto, este estudio se centrará en la sintaxis profunda, un subtipo de sintaxis de dependencias **{RSintP$_{k1}$}**, concebida como una serie de estructuras constituidas por actantes y adjuntos. En español puede haber un máximo de seis actantes (I, II, ..., VI) y de tres adjuntos: atributivo (ATTR), apenditivo (APPEND) y coordinativo (COORD). Según Mel'cuk (2003:7)[2]:

"The *Deep-Syntactic Structure* of a sentence is a tree whose nodes are labeled with the full lexemes of the sentence, such that there is a one-to-one correspondence between DSyntnodes and full lexemes; the arcs of this tree, called branches, are labeled with names of abstract universal Deep-Syntactic Relations".

Será con este tipo de sintaxis profunda de dependencias con la que trabajaremos en nuestro estudio, llevando a cabo el análisis de nuestros textos médicos en forma de estructura arbórea de dependencias, marcando los actantes de las formas verbales y cualquiera de los tres tipos de adjuntos que puedan aparecer (ATTR, APPEND, COORD). Este tipo de sintaxis tiene la ventaja, de cara a nuestros propósitos, de ser suficientemente detallada y de reflejar con exactitud las relaciones de dependencias

[2] Por otro lado nos encontramos con otro subtipo de sintaxis de dependencias, la superficial, que no trataremos en este estudio pero que nos parece necesario mencionar, en palabras también de Mel'cuk (2003:7): "The *Surface-Syntactic Structure* of a sentence is also a tree whose nodes are labeled with all the lexemes of the sentence (including all auxiliary and 'structural' words), again there being a one-to-one correspondence between the SSynt-nodes and the lexemes; the arcs of this tree, also called branches, are labeled with names of language-specific Surface-Syntactic Relations, each of which represents a particular construction of the language (their number, in an average language, is somewhere around 50)".

existentes en la lengua, utilizando un pequeño y limitado número de relaciones, al contrario de la sintaxis superficial, que utiliza muchas más. Además esta estructura es independiente de lengua, ya que se puede generalizar para cualquiera de ellas, y permite la captura de las relaciones más relevantes.

Debemos destacar también el énfasis que desde la *Meaning-Text Theory* (Mel'cuk, 2001) se hace de la estructura sintáctica comunicativa profunda (destacando la contraposición existente entre Tema-Rema) que, como veremos más adelante, también intentaremos explotar para intentar extraer nuevas informaciones de cara al resumen.

En el gráfico 3 observamos un fragmento de estructura arbórea de sintaxis de dependencias con un núcleo verbal del que dependen un APPEND, un Actante I y un Actante II de los cuales a su vez dependen otros elementos: "Así mismo, la mayor accesibilidad al hospital (residir en la misma ciudad) se asocia con un mayor número de visitas innecesarias al servicio de urgencias."

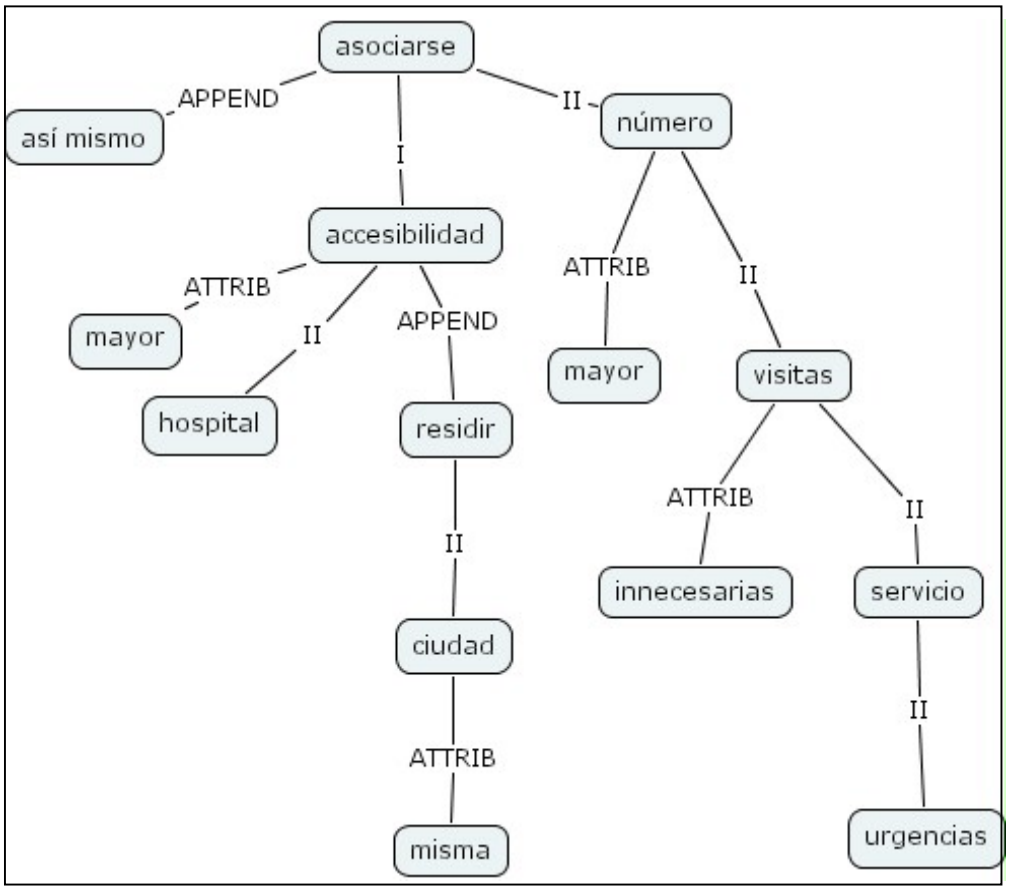

Gráfico 3. Fragmento de estructura arbórea de la sintaxis de dependencias

4.3. La estructura comunicativa

Los primeros estudios sobre la organización comunicativa de los textos comenzaron a finales del siglo XIX (Weil, 1844) y han ido evolucionado a lo largo del tiempo. Algunos trabajos relevantes son los de von der Gabelentz (1869), la Escuela de Praga (Sgall *et al*, 1973, 1986), Halliday (1967a, 1967b, 1967c, 1985), Kuno (1972), Vallduvi (1990) y Lambrecht (1994), entre muchos otros. En este estudio seguiremos la concepción de organización comunicativa de Mel'cuk (2001), basada en la estructura semántica, la estructura sintáctica y la estructura comunicativa.

Mel'cuk (2001) afirma que para obtener una completa representación de la estructura semántico-comunicativa se requieren al menos ocho oposiciones comunicativas: *Thematicity, Giveness, Focalization, Perspective, Emphasis, Presupposedness, Unitariness* y *Locucionallity*. Indica (Mel'cuk, 2001:18), además, que en la estructura comunicativa se distinguen dos elementos, el Predicado Comunicativo y el Sujeto Comunicativo:

"The Communicative Predicate is that part of the meaning of an utterance which is presented (by the Speaker) as being communicated. It is also called the Rheme, or Comment. The Communicative Subject is what the Rheme applies to and communicates about. It is called the Theme, or Topic."

Para nuestros propósitos de cara al desarrollo de un modelo de resumidor automático sólo será pertinente la primera de las ocho oposiciones, es decir, la contraposición entre Tema y Rema[3], que utilizaremos en el desarrollo de nuestras reglas sintáctico-discursivas. Esta elección se debe a que dicha oposición será más relevante de cara al resumen que las otras siete, porque nos indicará de qué se habla en cada oración y qué se dice acerca de ello, lo cual nos resultará de gran ayuda para discernir cuál será la información indispensable para el resumen automático

[3] En la oposición comunicativa de *Thematicity* también habría que considerar un tercer elemento, el *Sem-Com-Specifier,* que en este trabajo no tendremos en cuenta por no aportar informaciones imprescindibles de cara al resumen. Estos elementos pueden ser internos (Comm-Circumstantials y Comm-Characterizers) o externos (Comm-Connectors). En palabras de Mel'cuk (2001:96):
"From the viewpoint of the message-oriented organization of its propositional meaning, the Speaker first has to distinguish two major parts of S:
1) the **Communicative Core**, which constitutes the (logical) PROPOSITION carried by the utterance that expresses S [...];
2) and the remainder, which constitutes the set of **Communicative Specifiers**; as their name implies, Communicative Specifiers express some SPECIFICS concerning the Comm-Core."

En el Gráfico 4 observamos un ejemplo de análisis sintáctico profundo de dependencias con la superposición de la oposición entre Tema y Rema: "Los servicios de urgencias hospitalarios son cada vez más utilizados por la población".

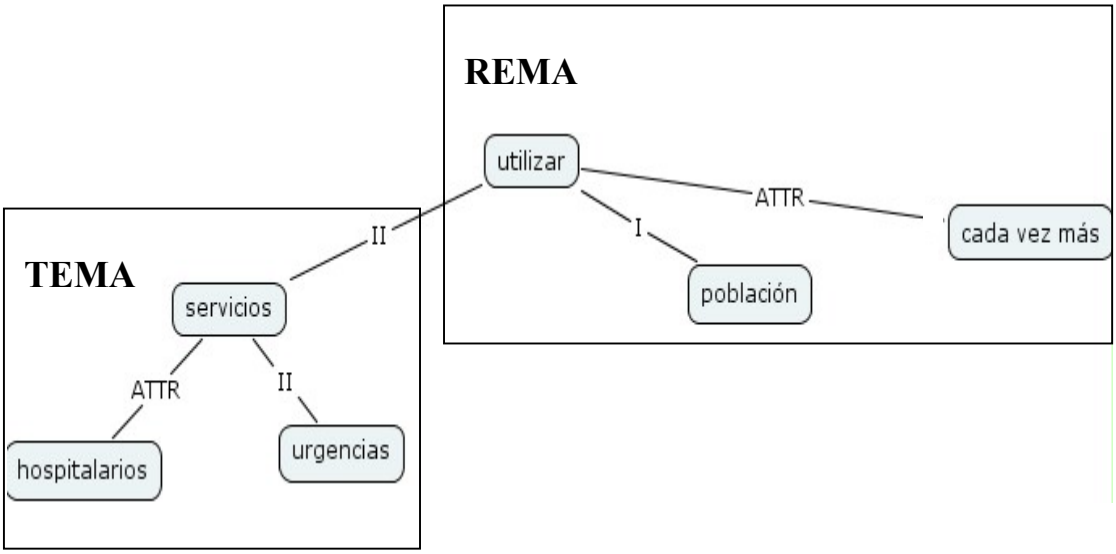

Gráfico 4. Fragmento de estructura arbórea de sintaxis profunda de dependencias y de estructura comunicativa (Tema-Rema)

5. Objetivos, hipótesis y objeto de estudio

5.1. Objetivos

El objetivo de la tesis es encontrar un modo de integración de las ventajas de varias perspectivas lingüísticas (textual, léxica, discursiva y sintáctico-comunicativa) orientado a la creación de un modelo de sistema de resumen automático. Habrá que concretar en qué ayuda cada perspectiva al futuro resumen, y cómo debemos unir estas informaciones para finalmente recopilar datos que nos permitan extraer conclusiones mediante las cuales crear un sistema de reglas que siente las bases de dicha tesis.

Es importante destacar que la implementación del modelo no es uno de nuestros objetivos, mientras que sí lo es el acercamiento lingüístico que llevaremos a cabo para el desarrollo de dicho modelo.

5.2. Hipótesis de partida

Nuestro trabajo se desarrolla sobre la base de las siguientes hipótesis:

1. La utilización de un solo tipo de información lingüística no es suficiente para llegar a un buen resumen.

2. Deben integrarse informaciones lingüísticas de varios tipos para poder llegar a una completa representación de los textos y a un posterior resumen de los mismos.

2.1. Debe utilizarse información textual, léxica, discursiva y sintáctico-comunicativa.

2.2. La simple utilización de la información discursiva no será lo suficientemente válida por sí sola para llegar a un resumen totalmente coherente y cohesivo.

2.2.1. Los supuestos de Marcu (2000) en cuanto a la relación entre los postulados de la RST y la generación de resúmenes son válidos, pero insuficientes en determinadas ocasiones en las que se necesite un análisis más profundo, ya que él se limita a utilizar la información discursiva.

2.2.2. Debe profundizarse a nivel sintáctico mediante la sintaxis de dependencias y la estructura comunicativa para poder llegar a una completa representación de la totalidad de un texto y, por tanto, a un posterior resumen del mismo.

2.3. Debe considerarse la competencia de los profesionales del dominio, que se refleja en unidades léxicas relevantes.

Observemos un caso que ilustre el punto 2.2. Según Marcu (2000) los fallos de su sistema de resumen automático basado en la estructura discursiva pueden deberse básicamente a dos motivos: o que el analizador discursivo no construya adecuadamente las estructuras arbóreas, o que el mapeo de las estructuras discursivas que da puntuaciones a sus elementos dependiendo de su importancia sea demasiado simple.

Marcu (2000) explica que la intuición de este acercamiento es que las unidades textuales que están en los conjuntos de promoción de los nodos superiores de un árbol discursivo son más importantes que las unidades que salen de los nodos encontrados en la parte de abajo. Un modo muy simple de inducir una clasificación es calculando una puntuación para cada unidad elemental del texto sobre la base de la profundidad del nodo de la estructura arbórea, donde la unidad que aparezca antes es una unidad de promoción. Cuanto mayor sea la puntuación de la unidad, más importante será considerada esa unidad dentro del texto. A veces, este acercamiento no es del todo fiable, y como el propio Marcu (2000:225) afirma:

"Mechanisms that are not inherent to the rhetorical structure of the text are needed in order to explain why one nucleus of a multinuclear relation is considered important by humans."

Observemos ahora un ejemplo concreto. Marcu (2000) indica que, en algunas ocasiones, para ciertos tipos de relaciones de *Elaboración* parece necesario asignar una puntuación mayor a sus satélites que la que actualmente ofrece su fórmula. Él ofrece un

ejemplo en el que su programa descarta dos elementos que los especialistas humanos de su experimento no dudan en seleccionar para el resumen manual.

Ej.1 "[Smart cards have two main advantages over magnetic-stripe-card.[3]][First, they can carry 10 or even 100 times as much information[4]] [–and hold it much more robustly.[5]][Second they can execute complex tasks in conjunction with a terminal.[6]]"

Ningún experto duda en seleccionar para el resumen las unidades 3, 4 y 6, mientras que el programa sólo selecciona la unidad 3, ya que a la 4 y la 6 se les da una puntuación baja por ser satélites de la unidad 3 en relación de *Elaboración*. Por tanto, habría que buscar estrategias que mejorasen este aspecto.

En nuestro corpus encontramos un caso parecido:

Ej.2 "[El análisis de regresión logística identificó tres variables asociadas, de forma independiente, con una visita apropiada a urgencias:[1]] [acudir a este servicio por indicación de un médico,[2]] [vivir fuera de la región respecto a residir en la ciudad en la que está el hospital[3]] [y pertenecer a los grupos de consultas quirúrgicas y traumatismos respecto a la enfermedad médica y pediatría.[4]]"

En el Ej.1 encontramos un primer actante sintáctico profundo que coincide con el Tema de la oración (I: *They*) y un segundo actante sintáctico profundo (II: *advantages*), seguidos por más detalles del actante I. En concreto, los dos satélites son una *Elaboración* del Rema del núcleo (que en este caso coincide con el actante II), y el Tema del núcleo (que en este caso coincide con el actante I) es el mismo que los Temas de los dos satélites. En el Ej.2 la situación es muy similar (I: análisis / II: variables).

Nuestra idea es que mediante la integración del discurso y la sintaxis pueden llegar a crearse una serie de reglas que solucionen el tipo de problemas que Marcu (2000) señala. En el apartado 7.2. veremos las que hemos desarrollado hasta ahora, incluyendo la que solucionaría el problema observado en el ejemplo de Marcu.

5.3. Objeto de estudio

El objeto de estudio de la tesis es la estructura textual, léxica, discursiva y sintáctico-comunicativa de artículos médicos en español (en concreto, extraídos de la

revista *Medicina Clínica* y que, a su vez, forman parte del subcorpus de medicina del Corpus Técnico del Instituto Universitario de Lingüística Aplicada[4]) de cara al desarrollo de un modelo de sistema de resumen automático.

[4] Este corpus puede consultarse en: http://brangaene.upf.es/bwananet/index.htm

6. Corpus de análisis

El corpus utilizado en este proyecto de tesis está formado por 20 artículos médicos en español con sus respectivos resúmenes extraídos de la revista *Medicina Clínica*. Esta revista, fundada en 1943, es la única publicación semanal de contenido clínico que se edita en España y 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, la actualidad de los temas y sobre todo su sentido práctico, buscando siempre que la información sea de la mayor utilidad en la práctica clínica. *Medicina Clínica* es un vehículo de información científica de reconocida calidad, como demuestra su inclusión en los más prestigiosos y selectivos índices bibliográficos del mundo: *Science Citation Index*, *Current Contents*, *Index Medicus* y *Excerpta Medica.*

A su vez, estos artículos forman parte del Corpus Técnico del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra de Barcelona. En concreto, se han extraído del subcorpus del ámbito de la medicina. 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).

7. Metodología de trabajo

La metodología que se seguirá en la tesis (dejando de lado de momento los criterios textuales y léxicos, que se explicarán más adelante) incluye varios pasos. En primer lugar se llevará a cabo un doble análisis manual de los textos de nuestro corpus: discursivo y sintáctico-comunicativo. En segundo lugar se observarán los resultados de dichos análisis e intentarán extraerse conclusiones de cara al resumen automático, que se explicitarán en forma de reglas sintáctico-discursivas. En tercer lugar se aplicarán dichas reglas a algunos textos seleccionados al azar (pero también de la revista *Medicina Clínica*) que formarán nuestro corpus de contraste, para comprobar la efectividad de las reglas desarrolladas. Finalmente se observará la coincidencia entre los contenidos de nuestros resultados y los del resumen del autor, para constatar si el resumen resultante de la aplicación de las reglas creadas se aproxima al máximo al del autor (considerando, como hemos comentado en el apartado 3.3, que el resumen del autor es el adecuado).

7.1. Análisis manual de los textos del corpus

7.1.1. Análisis discursivo de la RST

Como hemos explicado en el apartado 4.1, para el análisis discursivo de los textos de nuestro corpus partiremos de las relaciones de la *Rhetorical Structure Theory* de Mann & Thompson (1988). El listado completo de relaciones establecidas en un principio por los autores se encuentra en el Anexo 3. Por un lado, las relaciones más frecuentes son las que relacionan dos unidades de texto de tal manera que una de ellas tenga un papel específico con respecto a la otra, es decir, que una sea el núcleo (que aporta la información esencial) y otra sea su satélite (que aporta información adicional al núcleo). Ejemplos serían relaciones núcleo-satélite de *Evidencia*, *Fondo*, *Preparación* o *Concesión*. Por otro lado, hay relaciones que no presentan una unidad central con respecto a los propósitos del autor, denominadas *Multinucleares*. Ejemplos serían relaciones de *Contraste*, *Lista* o *Secuencia*.

Así, la metodología consistirá en marcar sobre los textos de nuestro corpus cada relación discursiva existente en ellos y crear árboles de estructuras discursivas a partir de dichas relaciones.

En el Anexo 4 pueden observarse los árboles de estructuras discursivas creados a partir uno de los artículos médicos que forman nuestro corpus: "Visitas inapropiadas al servicio de urgencias de un hospital general", en donde se ven reflejadas diversas relaciones, que permiten hacerse una idea general del funcionamiento de este planteamiento.

El análisis de las relaciones se ha llevado a cabo de manera manual, pero se ha utilizado como soporte informático la *RSTtool* diseñada por Marcu para que resulte más sencilla la visualización de los árboles de estructuras. La traducción que aparece entre paréntesis de cada relación discursiva (Anexo 3), se debe a que esta herramienta, aunque puede utilizarse para el análisis de textos de cualquier lengua, está diseñada para usar como interfaz la lengua inglesa.

7.1.2. Análisis sintáctico de dependencias

Como ya hemos comentado en el apartado 4.2, en este estudio se utilizará la sintaxis profunda de dependencias integrada en la *Meaning-Text Theory* (Mel'cuk, 1988), evidenciada mediante estructuras arbóreas.

La metodología que seguiremos consistirá en llevar a cabo el análisis de nuestro corpus de artículos médicos en forma de estructura arbórea de dependencias, marcando todos los actantes (en español puede haber un máximo de seis: I, II, ..., VI) y los adjuntos que aparezcan: atributivo (ATTR), apenditivo (APPEND) y coordinativo (COORD).

Tendremos en cuenta además la contraposición existente entre Tema y Rema, que en ocasiones ofrecerá información importante para al resumen, ya que la estructura comunicativa profunda integrada en la *Meaning-Text Theory* (Mel'cuk, 2001) será de importancia en casos que la simple utilización de actantes no pueda ofrecernos información suficiente para realizar generalizaciones.

En el Anexo 5 pueden observarse los árboles de estructuras sintácticas (profundas) de dependencias con sus correspondientes estructuras comunicativas creados a partir de uno de los artículos médicos que forman nuestro corpus: "Visitas inapropiadas al servicio de urgencias de un hospital general", en donde se ven reflejados los actantes y los adjuntos existentes, que permiten hacerse una idea general de como funciona este planteamiento.

La herramienta de la que nos hemos servido se llama *IHMC Cmap Tools*, una interfaz sencilla y eficaz que, en principio, está diseñada para la elaboración de árboles conceptuales, pero que nos permite reflejar con claridad cualquier tipo de elemento o relación deseada. Se han marcado todos los casos de correferencia con una línea punteada.

7.2. Desarrollo de las reglas sintáctico-discursivas: primeros resultados

Como ya hemos comentado anteriormente, la creación de las reglas sintáctico-discursivas que aplicaremos posteriormente sobre los textos que queramos resumir se llevará a cabo a partir de la integración de la perspectiva discursiva (mediante las relaciones de la RST) y sintáctico-comunicativa (mediante relaciones sintácticas profundas de dependencias y mediante la contraposición entre Tema y Rema).

Veamos un ejemplo que ilustre esta idea. La regla que solucionaría el problema propuesto por Marcu (2000) y que nosotros hemos reflejado en el apartado 5.2. sería la siguiente[5]:

IF *S* is satellite of ELABORATION E_1
and *S* elaborates on the Rheme of the nucleus *N* of E_1
and the Theme of *S* is equal to the Theme of *N* (or: **I** of *S* is equal to **I** of *N*)
THEN KEEP *S*

Si aplicásemos esta regla sobre los ejemplos reflejados en 5.2. obtendríamos la información adecuada para el resumen, siempre comparando nuestros resultados con los de expertos humanos.

El trabajo de compilación de las reglas está aún en curso y será fundamentalmente el grueso de la futura tesis, junto con la búsqueda de ejemplos y la experimentación para validar nuestras suposiciones teóricas. En este proyecto hemos creado ya algunas de ellas (y se ofrecen aquí con algunos de sus respectivos ejemplos

[5] Las reglas sintáctico-discursivas están reflejadas en inglés teniendo en cuenta que, ya que el proyecto de tesis y la futura tesis estarán escritos en español, los lectores que no sepan este idioma tengan más posibilidades de entenderlas.

extraídos de nuestro corpus de artículos médicos[6]), pero somos conscientes de que la mayoría de ellas aún están por desarrollar.

Reglas sintáctico-discursivas con algunos ejemplos representativos extraídos de artículos médicos de la revista *Medicina Clínica* provenientes del Corpus Técnico del IULA[7]

(1)
IF *S* is satellite of ELABORATION E_1
and *S* is ATTR of an element of the nucleus of E_1
THEN ELIMINATE *S*

Si encontramos un satélite de una relación de *Elaboración* que además sea un ATTR, éste puede ser eliminado. En muchas ocasiones estos fragmentos son oraciones de relativo explicativas (evidenciadas por una coma antes del "que" relativo o del "lo que", "lo cual", etc.).

Ej. I. Esta cifra es superior al 26,8% de Oterino *et al***,** que emplean, también, el PAUH para el análisis de las visitas inadecuadas a dicho servicio. (m00788)[8]

Ej. II. El objetivo de este estudio fue analizar la evolución de la prevalencia de infección por el VIH en las madres de los nacidos entre 1996 y 1999 en siete comunidades autónomas (CCAA), que representan el 26,5% de la población y el 25% de los nacimientos en España. (m00794)

Ej. III. Los resultados obtenidos, sobre todo los que se refieren a la relación entre la prevalencia de anticuerpos y el lugar de nacimiento o de residencia, deben valorarse con cautela, ya que se trata de un estudio retrospectivo y el muestreo no es el más adecuado para su análisis. (m00792)

[6] Algunos otros ejemplos pueden observarse en el Anexo 6.
[7] Los fragmentos de texto que aparecen subrayados podrían ser eliminados.
[8] La letra y el número que aparecen entre paréntesis después de cada ejemplo indican el texto del Corpus Técnico del que se ha extraído. La letra "m" explicita que se trata del ámbito de la medicina y el número es el código sgm.

(2)

IF *S* is satellite of EVALUATION or of INTERPRETATION of the nucleus *N*

and *S* is ATTR of an element of N

THEN ELIMINATE *S*

También podría aplicarse la regla anterior con satélites de *Evaluación* o *Interpretación*.

Ej. IV. La mitad de las visitas se justificó por la realización de pruebas complementarias, lo que indica que los servicios de urgencias extrahospitalarios no cubren por completo las necesidades de la población general. (m00788)

Ej. V. Coincidiendo con ese mismo estudio, la visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias, lo que estaría en relación con el papel de filtro de la atención primaria. (m00788)

Ej. VI. Desde 1995 se ha producido en España un descenso del 81% en los casos de sida por transmisión perinatal, lo que refleja un balance global favorable hasta el momento en esta vía de transmisión del VIH. (m00794)

(3)

IF *S* is satellite of ELABORATION E_1

and S is APPEND of an element of the nucleus *N* of E_1

THEN ELIMINATE *S*

Si encontramos un satélite de una relación de *Elaboración* que además sea un APPEND, éste puede ser eliminado. En muchas ocasiones estos fragmentos aparecen entre paréntesis.

Ej. VII. Además, se pidió a un médico adjunto de urgencias (con más de 5 años de experiencia) que no conocía el cuestionario PAUH que indicase, según su criterio clínico, cuáles de las 288 visitas eran inapropiadas. (m00788)

Ej. VIII. El algoritmo diagnóstico incluyó un cribado de todas las muestras con una prueba de ELISA (Genelavia Mixt HIV-1/2, Sanofi-Pasteur, Francia), un doble ELISA en paralelo para las reactivas o dudosas y la confirmación posterior mediante inmunoblot (Chiron RIBA HIV-1/2 SIA, Ortho Diagnostic, EE.UU.). (m00794)

Ej. IX. Cocer la carne adecuadamente (los quistes se inactivan a temperaturas superiores a 60°C) es una medida preventiva, al igual que lavar las frutas y las hortalizas (preferiblemente comer verduras cocidas), evitar el contacto con gatos que puedan cazar libremente y utilizar guantes al realizar trabajos de jardinería. (m00792)

(4)

IF *S* is satellite of ELABORATION E_1
and the nucleus *N* of E_1 contains Rheme *R*
and *S* contains *R* as Theme
THEN ELIMINATE *S*

Si encontramos una relación de *Elaboración* en la que el Rema del núcleo sea el Tema del satélite, dicho satélite puede ser eliminado. En los siguientes ejemplos dichos Rema y Tema aparecen en cursiva.

Ej. X. El PAUH no identificó *a dos niños cuya visita a urgencias fue considerada apropiada por el experto*; *ambos* eran lactantes.(m00788)

Ej. XI. De hecho, en los años 1998 y 1999 se diagnosticaron *cuatro casos de toxoplasmosis gestacional entre las mujeres atendidas por el departamento de obstetricia y ginecología de nuestro hospital. Una de ellas* fue diagnosticada al final del embarazo y dio a luz un niño con toxoplasmosis congénita grave. (m00792)

Ej. XII. No se han observado *diferencias en el período de incubación del sida en función del año de seroconversión*. *El efecto del año de seroconversión*, presente en el análisis bruto, desaparece cuando se ajusta por cada cohorte individual (cociente de verosimilitudes entre el modelo con y sin efecto, p = 0,3787). (m00793)

(5)

IF *S* is satellite of CONTRAST *C*

and the nucleus *N* of *C* contains a Rheme *R*

and *S* contains *R* as Theme

THEN ELIMINATE *S*

También podría aplicarse la regla anterior en relaciones *Multinucleares* de *Contraste*.

Ej. XIII. El análisis de regresión logística identificó *tres variables asociadas, de forma independiente, con una visita apropiada a urgencias: acudir a este servicio por indicación de un médico, vivir fuera de la región respecto a residir en la ciudad en la que está el hospital y pertenecer a los grupos de consultas quirúrgicas y traumatismos respecto a la enfermedad médica y pediatría. El resto de las variables* (edad, género, mes, hora y día de la visita a urgencias) no se asoció con el uso adecuado del servicio de urgencias hospitalario. (m00788)

Ej. XIV. *Cuarenta y ocho pacientes* (16,8%) fueron enviados por un médico al servicio de urgencias; *el resto* acudió por propia iniciativa. (m00788)

(6a)

IF CONTRAST between *N1* and *N2*

and *N1* and *N2* possess the same Theme

THEN KEEP *N1* and *N2*

Si encontramos una relación *Multinuclear* de *Contraste* en la que los dos núcleos de la relación posean el mismo Tema debemos mantener información de ambos para que no se pierda esa idea de contraste entre dos elementos. En el siguiente ejemplo el Tema aparece en cursiva.

Ej. XV. *Todos los recién nacidos de madres infectadas por el VIH* presentan anticuerpos anti-VIH recibidos de la madre, pero *muchos de estos niños* no están infectados y los pierden antes del año y medio. (m00794)

(6b)

IF CONTRAST between *N1* and *N2*

and *N1* and *N2* possess different Themes and different Rhemes

THEN KEEP *N1* and *N2*

Si encontramos una relación *Multinuclear* de *Contraste* en la que los dos núcleos de la relación posean diferentes Tema y Rema, debemos mantener información de ambos. En los siguientes ejemplos el Rema aparece en cursiva y el Tema en negrita.

Ej. XVI. **Baleares** destacó *con una seroprevalencia mayor que el resto*, próxima a las obtenidas en 1997 en Cataluña (2,3 por 1.000) y en 1998 en la Comunidad Valenciana (2,5 por 1.000, datos no publicados). Por el contrario, **otras CCAA** no alcanzaron *la mitad de estas cifras*. (m00794)

Ej. XVII. La prevalencia del **VIH-1** encontrada se corresponde *con una situación endémica*, mientras que el **VIH-2** presenta *un patrón de casos esporádicos*. (m00794)

(7)

IF LIST *L* of $N_1 \ldots N_n$

and *L* is a DSynt-actant

THEN

IF n > ⌊_

THEN eliminate $N_1 \ldots N_n$

ELSE keep $N_1 \ldots N_n$

Si encontramos una enumeración de actantes (bien sean I, II o III) que sean los elementos de una relación de *Lista*, no se eliminará ningún elemento o, por el contrario, si la lista es demasiado larga (habrá que decidir un límite ⌊_), se eliminará todo el fragmento en el que aparezcan.

Ej. XVIII. Los centros de reclutamiento fueron los tres Centros de Información y Prevención del Sida (CIPS) de la Comunidad Valenciana ubicados en Alicante, Castellón y Valencia; los Centros de Atención y Prevención del Sida (CAPS) de

Barcelona; el Centro Sanitario Sandoval (centro de prevención y tratamiento de enfermedades de transmisión sexual) de Madrid; las Unidades de Hemofilia del Hospital Vall d'Hebron de Barcelona y La Fe de Valencia, y la Unidad de VIH del Hospital Germans Trias i Pujol de Badalona. (m00793): **se elimina todo.**

Ej. XIX. Se definió como seroconversor al sujeto VIH positivo que cumpliera cualquiera de las condiciones siguientes:
1. Se dispone de un test para el VIH negativo previo.
2. Se demuestra una enfermedad aguda por el VIH con seroconversión (patrones de laboratorio característicos).
3. Se tiene evidencia objetiva y sólidamente documentada sobre cuándo se infectó el sujeto. En este grupo se incluyen los afectados de hemofilia y otros casos excepcionales (p. ej., persona VIH positiva con 16 años de edad que en los últimos 3 años ha tenido una única pareja sexual seropositiva al VIH y carece de otras prácticas de riesgo). (m00793): **se elimina todo.**

Ej. XX. El algoritmo diagnóstico incluyó un cribado de todas las muestras con una prueba de ELISA (Genelavia Mixt HIV-1/2, Sanofi-Pasteur, Francia), un doble ELISA en paralelo para las reactivas o dudosas y la confirmación posterior mediante inmunoblot (Chiron RIBA HIV-1/2 SIA, Ortho Diagnostic, EE.UU.). (m00794): **se conserva todo (menos los fragmentos entre paréntesis, que se eliminan después de la aplicación de la regla 3).**

(8)
IF SEQUENCE *S* of $N_1 \ldots N_n$
and *S* is a DSynt-actant
THEN
IF n > ⌊_
THEN eliminate $N_1 \ldots N_n$
ELSE keep $N_1 \ldots N_n$

La regla anterior también puede aplicarse con la relación *Multinuclear* de *Secuencia*.

Ej. XXI. Entre 1996 y 1999, se detrajeron sistemáticamente 1 o 2 gotas sobrantes de sangre de cada recién nacido sin ninguna identificación. Se almacenaron en bolsas cuya única referencia fue la provincia y el año de nacimiento, y se conservaron entre 4 y 8 ºC hasta su remisión trimestral al laboratorio de retrovirus del Centro Nacional de Microbiología. Las manchas de sangre fueron cortadas mediante un perforador semiautomático en discos de 3 mm de diámetro y eluidas en tampón fosfato salino-Tween, para la determinación de anticuerpos frente al VIH-1 y 2. (m00794): **se elimina todo.**

(9)

IF UNION of *N1* and *N2*

and *N1* and *N2* are DSynt-actants

THEN KEEP *N1* and *N2*

Una parte de la regla anterior también puede aplicarse a la relación *Multinuclear* de *Unión*.

Ej. XXII. De los 1.054 sujetos, 800 cumplían la condición 1, **y** ninguno cumplía la condición 2. (m00793)

Ej. XXIII. El seguimiento de los seroconversores ha sido exhaustivo en cada uno de los centros, **y** habitualmente su información se actualiza cada 6 a 12 meses. (m00793)

Ej. XXIV. El seguimiento de estos pacientes, así como la incorporación de otros grupos de seroconversores, permitirá un conocimiento más profundo de la historia natural del VIH en nuestro país, **y** ayudará a evaluar el impacto de las nuevas terapias en el ámbito poblacional. (m00793)

(10)

IF *S* is satellite of CONCESSION *C*

and between the nucleus *N* of *C* and *S* the Dsynt relation COORD holds

THEN ELIMINATE *S*

Si encontramos un satélite de *Concesión* que sea a su vez uno de los elementos de una relación de Coordinación (sintáctica) éste puede ser eliminado para el resumen. En muchas ocasiones aparecen marcadores específicos de esta relación, que en nuestros ejemplos están marcados en negrita.

Ej. XXV. **A pesar de** la idea generalizada de que hay una gran sobreutilización de los servicios de urgencias, en este estudio, dos terceras partes de los pacientes acudieron a ellos de forma apropiada. (m00788)

Ej. XXVI. La seroprevalencia del conjunto de las CCAA aumentó un 56% entre 1996 y 1999, **si bien** el ascenso sólo se constató en Canarias, Castilla y León, y Castilla-La Mancha, las tres CCAA que presentaron menores seroprevalencias en 1996. (m00794)

Ej. XXVII. Como refleja la figura 1, se observa una tendencia decreciente de infección a lo largo del tiempo ($p < 0{,}001$). **Sin embargo**, existe un aumento de anticuerpos estadísticamente significativo entre el grupo de mujeres de 15 a 24 años y el de 35 a 45 en todos los años estudiados (excepto en 1999). (m00792)

Ej. XXVIII. El estudio demuestra que, **aunque** en conjunto la prevalencia de la infección por T. gondii está decreciendo entre las mujeres en edad fértil atendidas en nuestro hospital, existe una alta incidencia de primoinfección en este grupo de edad. (m00792)

(11)
IF *S* is satellite of BACKGROUND *B*
and the Theme of *S* is the Rheme of the nucleus *N* of *B*
THEN ELIMINATE *S*

Cuando encontremos un satélite de *Preparación* al principio de uno de los apartados que tenga un Tema que en el núcleo sea Rema, puede eliminarse dicho satélite para el resumen. En el siguiente ejemplo dicho elemento aparece en cursiva. Debemos tener en cuenta que el resumen al que nos proponemos llegar está destinado a especialistas en la

materia, que ya conocen seguramente los antecedentes de lo que se explica y a los que, en principio, no interesarán esos datos.

Ej. XXIX *La RM-mielografía* es una secuencia particular de RM que permite obtener imágenes del líquido cefalorraquídeo (LCR) en el interior del saco dural. Esta secuencia se caracteriza tanto por la rapidez en su obtención como por ser incruenta al no necesitar de la administración intratecal de un medio de contraste. Su principio técnico se basa en la obtención de imágenes extremadamente potenciadas en tiempos de relajación T2, lo que permite la visualización del LCR en el saco dural dentro del conducto espinal (por su T2 muy largo), con una supresión prácticamente completa de la señal del tejido sólido circundante.
Nuestro objetivo es conocer la rentabilidad de la *RM-mielografía* como técnica diagnóstica complementaria a la RM convencional en la valoración de las enfermedades de la columna vertebral. (m00795)

Ej. XXX. *La fractura vertebral* se ha considerado la más frecuente, y por ello es la más evaluada como criterio de selección y resultado en los ensayos clínicos terapéuticos. No obstante, comparada con las fracturas de cadera y antebrazo, sus estudios epidemiológicos son más limitados y tardíos.
El objetivo de este trabajo ha sido conocer la prevalencia de *la fractura vertebral* en nuestra población empleando los criterios radiológicos más usados en la actualidad para su diagnóstico siguiendo las recomendaciones del grupo de trabajo creado por la Fundación Americana de Osteoporosis. (m00796)

(12)
IF *S* is satellite of PURPOSE *P*
and between the nucleus *N* of *B* and *S* the Dsynt relation COORD holds
THEN ELIMINATE *S*

Si encontramos un satélite de *Propósito* que sea a su vez uno de los elementos de una relación de Coordinación (sintáctica) éste puede ser eliminado para el resumen. En muchas ocasiones aparecen marcadores específicos de esta relación, que en nuestros ejemplos están marcados en negrita.

Ej. XXXI. Los diferentes grupos que están trabajando con seroconversores en España se han articulado como Grupo Español Multicéntrico para el Estudio de Seroconversores (GEMES), **con el fin** de desarrollar líneas de trabajo conjunta. (m00793)

Ej. XXXII. **Con el fin** de minimizar las pérdidas en el seguimiento e identificar episodios de interés, los datos se han cruzado con diferentes registros, como el Registro Nacional de Sida, registros de información hospitalaria (CMBD), centros de desintoxicación, unidades de hospitalización a domicilio, prisiones en la red de Cataluña y los registros de mortalidad. (m00793)

Ej. XXXIII. **Para** evaluar el impacto de las nuevas terapias en el período de incubación del sida, es necesario conocer cuál era la situación previa a su introducción. (m00793)

7.3. Aplicación de las reglas sobre el corpus de contraste: nuestro resumen

El siguiente paso de nuestra metodología de trabajo será la aplicación de las reglas desarrolladas en los textos del corpus de contraste para observar los resúmenes resultantes.

A continuación, después de observar las reglas desarrolladas hasta ahora, veamos algún ejemplo claro en el que se pongan en práctica. Consideremos la aplicación de las siguientes dos reglas sobre un fragmento de texto de nuestro corpus que se ofrece a continuación:

(1)
IF *S* is satellite of ELABORATION E_1
and *S* is ATTR of an element of the nucleus of E_1
THEN ELIMINATE *S*

(3)
IF *S* is satellite of ELABORATION E_1
and S is APPEND of an element of the nucleus *N* of E_1
THEN ELIMINATE *S*

Ej. 3 “[Coincidiendo con ese mismo estudio,] [la visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias,] [lo que estaría en relación con el papel de filtro de la atención primaria.]” (m00788)

Mediante la aplicación de (1) y (3) la oración se reduce de la siguiente manera, coincidiendo con la información que el autor ha seleccionado para su resumen.

Ej. 4 “La visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias.” (m00788)

7.4. Comparación de nuestro resumen con el del autor

El siguiente y último paso de nuestra metodología de trabajo consiste en la validación de los resúmenes obtenidos a partir de la aplicación de nuestros criterios. Esta validación se llevará a cabo mediante la comparación de nuestro resumen con el del autor, como hemos comentado en el apartado 3.3.

En el apartado 9 observaremos un pequeño experimento en el que se evalúan los resúmenes resultantes de la aplicación de los criterios de los que disponemos hasta ahora, confirmándose la viabilidad de nuestro planteamiento.

8. Hacia un resumen justificado lingüísticamente

Como hemos venido observando, parece posible que mediante la integración de varias perspectivas lingüísticas (textual, léxica, discursiva y sintáctico-comunicativa) pueda llegarse a un modelo válido de generación automática de resúmenes. El Gráfico 5 muestra la arquitectura del sistema, que explicaremos a continuación en detalle.

8.1. Criterios para el resumidor

8.1.1. Criterios textuales

El primer paso para alcanzar el resumen deseado será la aplicación de los criterios textuales explicados en el apartado 3.1.2. sobre los artículos médicos de nuestro corpus, es decir, que se seleccionarán contenidos de cada uno de sus cuatro apartados, que siguen la estructura IMRD. Para la diferenciación de los apartados el modelo se basará en el reconocimiento de sus títulos, los cuales pueden presentar variaciones. Los títulos de los que disponemos hasta ahora están explicitados en la Tabla 4 del apartado 3.2.1, aunque en la tesis habrá que ampliar este estudio de reconocimiento de variantes de los títulos con más textos.

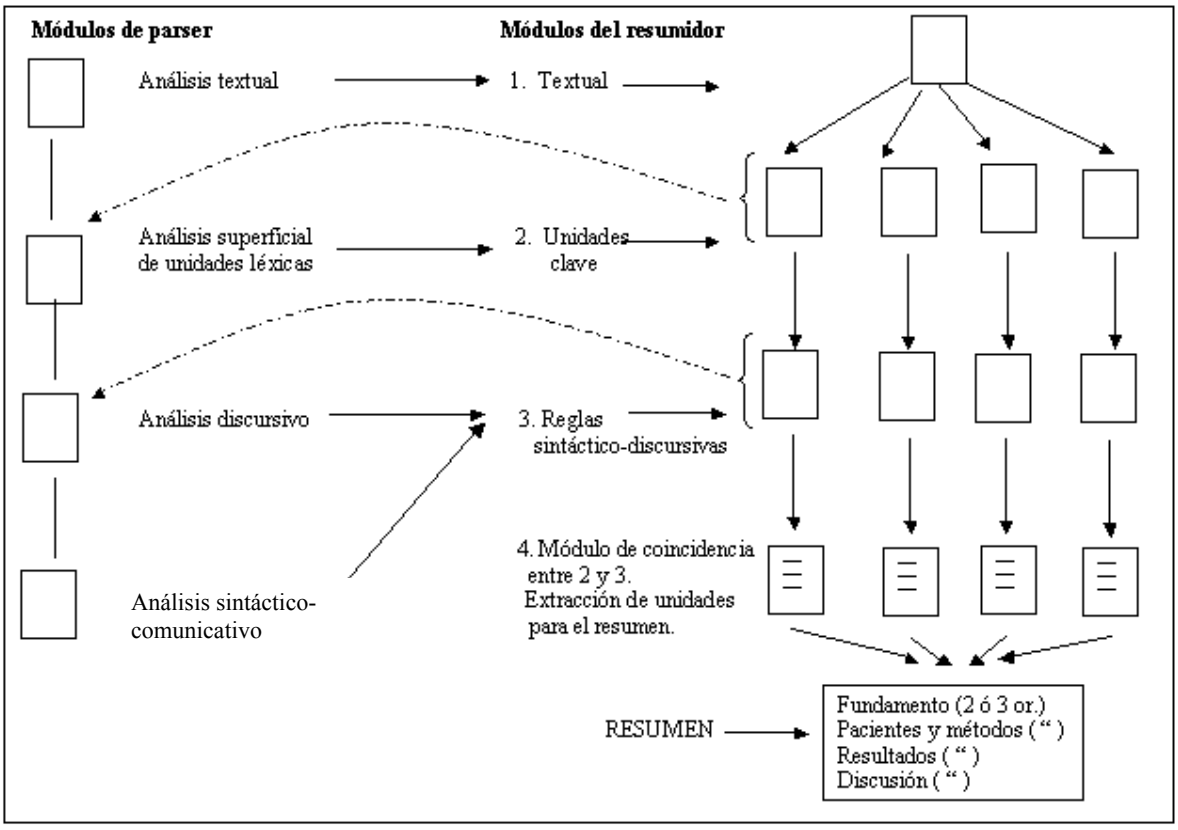

Gráfico 5. Arquitectura del sistema

8.1.2. Criterios léxicos: unidades representativas

Como se ha explicado anteriormente, consideramos que ciertas unidades léxicas pueden ser indicadores de relevancia en el tipo de textos sobre los que estamos trabajando (artículo médico). Después de la primera aproximación realizada, se ha obtenido una lista provisional de unidades que se buscarán en cada uno de los cuatro apartados de los artículos. Una vez extraídos los fragmentos que contienen estas unidades léxicas se almacenarán para, posteriormente, observar la coincidencia entre estos criterios y los sintáctico-discursivos.

Para la comprobación de la relevancia de las unidades léxicas nominales se han realizado búsquedas con Bwananet, llegando a la conclusión de que, efectivamente, estas unidades reflejan informaciones importantes, siempre tomando como criterio de relevancia la coincidencia con los contenidos incluidos en el resumen del autor.

Para determinar qué verbos reflejan informaciones importantes se ha llevado a cabo un estudio orientativo sobre la frecuencia de los mismos, llegando a determinar 15 de ellos que en principio son los mejores indicadores de relevancia. Se han analizado los verbos de los resúmenes correspondientes a los 20 artículos médicos de nuestro corpus, pertenecientes a la revista *Medicina Clínica*. Se ha realizado un estudio de la frecuencia de aparición de los verbos más representativos de estos resúmenes, eliminado para el cómputo de la frecuencia todas las formas verbales de verbos auxiliares del tipo *ser*, *estar*, *haber*, *tener*, etc., ya que no se consideran representativos de esta área, ni puede afirmarse que con su utilización quiera marcarse alguna información relevante. Así, se han seleccionado los 15 verbos con aparición más frecuente. En un único caso nos encontraríamos con una locución verbal: *llevar a cabo*, como sinónimo de *realizar*. Los 15 verbos más frecuentes de este corpus de 20 resúmenes se observan en la Tabla 5.

Con la finalidad de comprobar si las oraciones que contienen estos verbos reflejan informaciones relevantes, se ha llevado a cabo una pequeña prueba. En primer lugar, se ha seleccionado al azar un artículo que no forma parte de nuestro corpus de análisis pero que se incluye en la misma revista, *Medicina Clínica*, titulado *Complicaciones en transportadores intestinales de paquetes con cocaína. Estudio de 215 casos*. En segundo lugar se han buscado en él todas las oraciones en las que apareciesen estos verbos. Finalmente se han comparado las oraciones seleccionadas con las oraciones del resumen que el mismo autor del artículo ha redactado. Los resultados

son positivos, ya que se observa que, de las 9 informaciones relevantes del resumen del autor, 6 aparecen reflejadas mediante 4 de los 15 verbos de nuestra lista.

VERBO	**FRECUENCIA** (nº de ocurrencias en el corpus de resúmenes)
Realizar	14
Asociar	12
Analizar	10
Presentar	10
Relacionar	6
Evaluar	9
Aportar	5
Estudiar	5
Valorar	4
Incluir	4
Observar	3
Llevar a cabo	3
Obtener	3
Alcanzar	2
Encontrar	2

Tabla 5. Número de ocurrencias de los 15 verbos más frecuentes en el corpus de 20 resúmenes

Este estudio deberá ampliarse en la tesis ya que observamos que, en ocasiones, los verbos seleccionados se encuentran en oraciones que no son utilizadas por el autor para su resumen. Se observa además una tendencia de aparición de mayor densidad léxica de las unidades de nuestra lista (tanto nominales como verbales) en los fragmentos que los autores seleccionan para sus resúmenes.

Las unidades léxicas obtenidas hasta ahora son las siguientes:

- Unidades nominales: *objetivo, objeto, propósito, intención, resumen, conclusión, resultado, estudio, trabajo*.
- Unidades verbales: *realizar, asociar, analizar, presentar, evaluar, relacionar, aportar, estudiar, valorar, incluir, observar, llevar a cabo, obtener, alcanzar, encontrar*.

8.1.3. Criterios sintáctico-discursivos

El tercer paso para llegar al resumen será la aplicación de los criterios sintáctico-discursivos en forma de las reglas reflejadas en el apartado 7.2. Como ya hemos aclarado anteriormente estas reglas deben ser ampliadas en la tesis para llegar a un modelo totalmente válido.

8.1.4. Coincidencia entre criterios léxicos y sintáctico-discursivos

Finalmente, el último paso para seleccionar los contenidos que se incluirán en el resumen será la comprobación de la coincidencia de los criterios léxicos y los criterios sintáctico-discursivos.

En los casos en que los dos tipos de criterios seleccionen los mismos contenidos, estos serán incluidos en el resumen sin ninguna duda. En los casos que no coincidan se tomarán los contenidos seleccionados por los criterios sintáctico-discursivos, ya que consideramos que estos son los que mejor muestran la evolución discursiva y sintáctica de los textos, y por tanto serán más fiables que los criterios léxicos por sí solos. Además, como hemos visto, en muchas ocasiones los elementos léxicos de nuestra lista aparecen en fragmentos que no son seleccionados para el resumen por el autor del artículo, con lo cual no podemos fiarnos exclusivamente de sus resultados.

El resumen resultante será pues el resultado de la aplicación de todos los criterios mencionados: textuales, léxicos, discursivos y sintáctico-comunicativos. Habrá que decidir la longitud deseada para el resumen, pero esto ya es un trabajo posterior de toma de decisiones cara la tesis.

8.2. Elementos para el análisis

Para la posible implementación de nuestro modelo de resumidor automático se necesitan varios elementos que lleven a cabo un análisis previo de los textos que se quieran resumir. Estos elementos serían el *parser* morfosintáctico (tokenizador, lematizador, desambiguador, analizador sintáctico), *parser* discursivo y *parser* comunicativo. Para ser más concretos, en este apartado nos referiremos a los recursos de

los que disponemos, sobre todo en el marco del Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra (UPF) de Barcelona.

Para una explicación completa del funcionamiento de las herramientas y una visualización de su arquitectura puede acudirse a Badia *et al.* (1998).

8.2.1. *Parser* morfosintáctico

8.2.1.1. Tokenizador

El tokenizador que podemos emplear forma parte del preproceso al que se someten los textos del Corpus Técnico del IULA, que incluye además detección de nombres propios, de abreviaturas, de frases,... Este tokenizador divide el texto en palabras claramente diferenciadas.

8.2.1.2. Lematizador

El *PALIC* (*Programa de Atribución de Lemas y Categorías*) es un analizador morfológico flexivo multilingüe que se aplica en la lematización y etiquetaje gramatical de los textos en castellano del Corpus Técnico del IULA (Yzaguirre *et al*, 2001a). En principio se aplica *PALIC* cuando los textos ya han sido marcados estructuralmente y preprocesados, pero también se ofrece la posibilidad de analizar textos que no han sido preparados para este corpus, lo cual amplia su ámbito de aplicación.

Ofrece de cada palabra (dividida por el tokenizador anterior) sus lemas y la categoría asociada a cada una de ellas:

forma X ---> lema 1 ----> cat 1

forma Y ---> lema 2 ----> cat 2

forma N ---> lema n ----> cat n

8.2.1.3. Desambiguador

Por un lado, *AMBILIC* es un programa de desambiguación de base lexicomorfológica para el catalán y el castellano que se utiliza en el tratamiento del Corpus Técnico del IULA (Yzaguirre *et al*, 2001b), después de las fases de marcaje

estructural, preproceso y análisis morfológico. Se aplica a textos analizados morfológicamente con el *PALIC*, con la finalidad de eliminar el máximo número de ambigüedades con reglas únicamente lingüísticas y por tanto con el máximo de fiabilidad.

Por otro lado, *MULTEX* es un proyecto europeo para el desarrollo de herramientas para el Procesamiento del Lenguaje Natural (PLN). En el marco de este proyecto se ha creado un desambiguador estadístico por ISSCO[9]. Esta herramienta se utiliza en el marco del Corpus Técnico del IULA y elimina mediante tratamiento estadístico las ambigüedades que no han podido solucionarse con *AMBILIC*.

8.2.1.4. Analizador sintáctico de dependencias

En el IULA está en curso el desarrollo de un analizador sintáctico de alto nivel lingüístico (texto etiquetado con información sintáctico-semántica) basado en el formalismo de unificación HPSG (Pollard & Sag, 1994) por parte de Marimón[10]. Este analizador utiliza la plataforma *Linguistic Knowledge Building* (LKB) (Copestake, 2002) desarrollado por las Universidades de Cambridge y Stanford. Cabría la posibilidad de convertir el resultado del analizador a relaciones de dependencias.

Además, en el IULA se está investigando la posibilidad de derivar una gramática de dependencias española de una gramática italiana que actualmente se utiliza en un analizador del italiano del *Istituto di Linguistica Computazionale* de la Universidad de Pisa (Lenci *et al*, 2003).

8.2.2. *Parser* discursivo

El analizador discursivo (basado en las relaciones de la RST) existe para el inglés (Marcu, 2000), y para el portugués hay un proyecto vigente en la Universidad de

[9] ISSCO, originalmente "Dalle Molle Institute for Semantic and Cognitive Studies", es ahora un grupo de investigación que forma parte de la School of Translation and Interpretation (ETI) / Multilingual Information Processing Unit (TIM) en la University of Geneva. Trabajan en el campo de Procesamiento del Lenguaje Natural (PLN).

[10] Montserrat Marimón es investigadora Juan de la Cierva del Instituto Universitario de Lingüística Aplicada (IULA-UPF) de Barcelona y realiza su investigación en el marco del proyecto "TEXTERM II: Fundamentos, estrategias y herramientas para el procesamiento y extracción automáticos de información" financiado por el Ministerio de Educación y Cultura español.

São Paulo denominado "DiZer" (Pardo, 2004) que pretende desarrollar este *parser* discursivo con relaciones de la RST para el portugués de Brasil.

De momento este analizador discursivo no existe para el español, lo cual será una limitación a la hora de implementar nuestro modelo de sistema de resumen automático. No existe en la actualidad ningún proyecto de investigación ni proyecto de tesis en España ni fuera de ella que se dedique a este tema para el español.

8.2.3. *Parser* comunicativo

No existe ningún *parser* comunicativo automático para el español, ni conocemos ningún proyecto que se dedique a este tema de investigación para esta lengua. En esta línea pero para el inglés y el checo encontramos los trabajos de Hajicova (1985, 1995) sobre detección automática de *Topic-Focus*.

9. Evaluación del estado actual: resultados preliminares

Para la evaluación de nuestro modelo de resumidor compararemos los resúmenes redactados por el mismo autor del artículo médico y los obtenidos a partir de la aplicación de nuestros criterios. Éste será el planteamiento que seguiremos para la evaluación del sistema en la tesis ya que, como se ha explicado en el capítulo 3.3, consideramos el resumen del autor como ideal en este ámbito.

Por el momento hemos aplicado este criterio en un pequeño experimento inicial que hemos llevado a cabo para observar la fiabilidad obtenida hasta ahora en nuestros resúmenes. Se han seleccionado cuatro artículos médicos de la revista *Medicina Clínica* como corpus de contraste y se han aplicado sobre ellos nuestros criterios léxicos y las reglas sintáctico-discursivas de las que disponemos hasta el momento (siendo conscientes de que es preciso ampliarlas y refinarlas). En concreto, se han aplicado sobre el apartado de *Introducción* de los cuatro textos, por tanto, al tratarse de una prueba para un apartado concreto, no cabe la posibilidad de aplicar nuestros criterios textuales.

Después de la aplicación de las reglas sintáctico-discursivas se observa que éstas han seleccionado positivamente los contenidos más relevantes de este apartado (siempre comparándolos con los del resumen del autor) que se corresponden en su mayoría con informaciones acerca del objetivo del artículo. En 3 de 4 resúmenes se selecciona correctamente la información más relevante del artículo.

En todos los fragmentos seleccionados por los criterios sintáctico-discursivos se encuentran unidades léxicas de nuestra lista. De todas maneras debemos tener en cuenta que en otros fragmentos no seleccionados por dichos criterios también aparecen algunas de estas unidades léxicas. Se observa, además, que las oraciones en las que coinciden los dos criterios hay más densidad de unidades léxicas que en los demás. Estos aspectos deberán ser estudiados con detalle en la futura tesis.

En el Anexo 7 puede observarse el apartado de Introducción de los cuatro textos de nuestro corpus de contraste con sus respectivos resúmenes creados tanto por el autor como a partir de nuestros criterios. En los textos 1, 2 y 3 nuestro resumen coincide con el del autor, mientras que en el texto 4 no lo hace.

A continuación veremos con detalle uno de los textos en los que la información relevante se ha seleccionado correctamente a partir de nuestros criterios. Las unidades léxicas que, como ya hemos explicado, consideramos relevantes aparecen subrayadas.

1. Complicaciones en transportadores intestinales de paquetes con cocaína. Estudio de 215 casos.

Introducción:

A los portadores de cuerpos extraños intraabdominales que contienen cocaína, con fines de contrabando, se les conoce con el síndrome del body packer. Los principales problemas médicos que se plantean en estos pacientes son: la sobredosificación de drogas por la rotura de uno de los paquetes y la obstrucción intestinal por impactación de dichos paquetes en el tubo digestivo.

Hemos estudiado la aparición de complicaciones en el seguimiento de individuos que ingieren estos paquetes de droga, con el fin de poder dar unas normas de actuación en estos casos.

El primer párrafo es un satélite de *Preparación* cuyo Tema "portadores de cuerpos extraños intraabdominales que contienen cocaína" se convierte en Rema en el núcleo "complicaciones en el seguimiento de individuos que ingieren estos paquetes de droga" con lo cual, después de la aplicación de la regla 11, será eliminado para el resumen:

(11)
IF *S* is satellite of BACKGROUND *B*
and the Theme of *S* is the Rheme of the nucleus *N* of *B*
THEN ELIMINATE *S*

A su vez, la segunda oración (dentro del primer párrafo y formando parte del Bakground) es un satélite de *Elaboración* de la primera, y además es un ATTR de "portadores" así que, después de la aplicación de la regla 1, será eliminada para el resumen:

(1)
IF *S* is satellite of ELABORATION E_1
and *S* is ATTR of an element of the nucleus of E_1
THEN ELIMINATE *S*

Finalmente, el segundo párrafo del texto está formado por un núcleo del que depende un satélite de *Propósito* que, después de la aplicación de la regla 12, será eliminado para el resumen:

(12)
IF *S* is satellite of PURPOSE *P*
and between the nucleus *N* of *B* and *S* the Dsynt relation COORD holds
THEN ELIMINATE *S*

Después de la aplicación de las reglas anteriores nuestro resumen sería el siguiente:

"Hemos estudiado la aparición de complicaciones en el seguimiento de individuos que ingieren estos paquetes de droga."

Como podemos observar, la información reflejada en nuestro resumen coincide con la del resumen del autor:

"Analizar las complicaciones aparecidas en las personas transportadoras de paquetes de cocaína."

Además, en el fragmento seleccionado después de la aplicación de las reglas sintáctico-discursivas se encuentra una de las unidades léxicas de nuestra lista de unidades representativas ("Hemos estudiado") así que, como coinciden ambos criterios, no hay duda de que la información seleccionada para el resumen es la adecuada.

10. Conclusiones

Como hemos visto, la idea de utilizar información textual, léxica, discursiva y sintáctico-comunicativa para resumir artículos médicos en español tiene posibilidades reales de aplicación cara a un modelo de sistema de resumen automático.

A lo largo de este proyecto de tesis se ha observado una panorámica general de la situación actual del resumen automático, se ha llevado a cabo un análisis del artículo médico como género, se han perfilado los dos marcos teóricos desde los que partimos (RST y MTT), y se ha explicado la metodología de trabajo que seguiremos, ofreciéndose unos resultados preliminares. Esta metodología consiste en la aplicación de criterios textuales (estructura IMRD), léxicos (selección de unidades nominales y verbales representativas), discursivos y sintáctico-comunicativos (en forma de reglas sintáctico-discursivas). Las reglas desarrolladas hasta el momento no constituyen un trabajo definitivo y concluyente, ya que deben ser contrastadas con más ejemplos y textos, pero forman una primera visión para observar cómo la integración de la estructura discursiva y de la estructura sintáctica y comunicativa puede ser un buen camino a la hora de buscar estrategias que solventen dificultades que parten del trabajo de Marcu (2000) y nos lleven a la mejora de los actuales sistemas de generación automática de resúmenes. Queda mucho trabajo por delante después de este trabajo exploratorio que no es un análisis exhaustivo, sino más bien un intento de configurar una primera visión que nos permita determinar las posibilidades de integración de varias perspectivas: textual, léxica, discursiva, y sintáctico-comunicativa. Con la unión de las ventajas que nos ofrezca cada visión parece que podrá llegarse a una correcta representación de la estructura de los textos, y a partir de aquí extraer conclusiones que nos aporten información relevante acerca de lo que es realmente importante en un documento y de lo que sería prescindible de cara a la generación de resúmenes automáticos.

Hasta ahora no hemos considerado las unidades léxicas clave con toda la profundidad que se necesita, pero ésta será una de las tareas que se desarrollará de cara a la tesis.

11. Trabajo futuro y plan de trabajo de la tesis

A partir de la defensa del proyecto de tesis doctoral se pretende seguir trabajando en esta misma línea e ir presentando los avances de nuestra investigación en congresos, tanto nacionales como internacionales, que favorezcan el intercambio de ideas sobre el tema.

El calendario que se propone de cara a la tesis es el siguiente:

OBJETIVO	FECHAS
Ampliación del experimento estadístico con los médicos (*Multidimensional Scaling*)	septiembre 2005 diciembre 2005
Ampliación del corpus: introducción de los textos en el Corpus Técnico del IULA (Bwananet)	septiembre 2005 diciembre 2005
Ampliación del estudio sobre subtítulos	enero 2006
Ampliación del estudio sobre unidades léxicas representativas	febrero 2006
Desarrollo de la totalidad de reglas sintáctico-discursivas para el *extract*	marzo 2006 junio 2006
Desarrollo de reglas de paráfrasis para el *abstract*	julio 2006 octubre 2006
Propuestas estadísticas de evaluación del sistema de resumen	noviembre 2006 febrero 2007
Comprobación de la fiabilidad de las reglas sintáctico-discursivas en otros ámbitos del Corpus Técnico	marzo 2007 mayo 2007
Redacción de la tesis doctoral	junio 2007 noviembre 2007
Defensa de la tesis doctoral	diciembre 2007

Título provisional de la tesis doctoral: "Hacia un modelo lingüístico de resumen automático de artículos médicos en español".

Bibliografía

ALONSO, L (2005). *Representing discourse for automatic text summarization via shallow NLP.* Tesis doctoral. Barcelona: Universidad de Barcelona.

ALONSO, L.; CASTELLÓN, I.; CLIMENT, S.; FUENTES, M.; PADRÓ, LL.; RODRÍGUEZ, H. (2003a). "Approaches to Text Summarization: Questions and Answers". *Revista Iberoamericana de Inteligencia Artificial* 22. 79-102.

ALONSO, L.; FUENTES, M. (2003b). "Integrating cohesion and coherence for Automatic Summarization". Actas de la EACL'03 Student Session. Budapest: ACL. 1-8.

ALONSO, L.; FUENTES, M. (2002). "Collaborating Discourse for Text Summarization". Actas de la Seventh ESSLLI'02 Student Session. Trento, Italy: ESSLLI.

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

BACH, C. (2005). "Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado". *Debate Terminológico* 1. París: RITERM (Red Iberoamericana de Terminología).

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

BADIA, T.; CABRÉ, T.; PUJOL, M.; TUELLS, A.; VIVALDI, J.; DE YZAGUIRRE, LL. (1998). "IULA's LSP Multilingual Corpus: compilation and processing". Actas de la ELRA conference. Granada. 29-31.

BARZILAY, R.; ELHAHAD, M. (1997). "Using lexical chains for text summarization". Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 10-17.

BÉLANGER, P.; KITTREDGE, R. (2005). "Paraphrasing with Space Constraints: Linguistic Operations in Journal Abstracting". Actas de la 2nd International Conference on the Meaning-Text Theory. Moscú: MTT.

BERGER, A.; MITTAL, V. (2000). "A system for summarizing Web Pages". Actas de la 23rd Annual Conference on Research and Development in Information Retrieval. Atenas: ACM. 144-151.

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

BLOOMFIELD, L. (1933). *Language*. New York: Holt, Rinehart & Winston.

BOGURAEV, B.; KENNEDY, C. (1997). "Salience-based content characterization of text documents." Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 2-9.

BRANDOW, R.; MITZE, K.; RAU, L. (1994). "Automatic condensation of electronic publications by sentence selection". *Information Processing and Management* 31. 675-685.

BRIZ, A. (2003). "Diccionario de partículas discursivas del español. Los resultados de un proyecto de investigación". Actas de la Sociedad Española de Lingüística: XXIII Simposio. Gerona: Univ. de Gerona.

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

CABRÉ, M. T. (2002). "Análisis textual y terminología, factores de activación de la competencia cognitiva en la traducción". En Alcina & Gamero (2002): *La traducción científico-técnica y la terminología en la sociedad de información*. Castellón: Univ. Jaume I. 87-105.

CARLSON, L. & MARCU, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISITR-545.

CASTEL, V. M.; AGUADO, L.; BOCCIA, C.; DIBLASI, A.; GARCÍA BRIZUELA, A.; HANSEN, A.; HASSAN, S.; HLAVACKA, L.; MIRET, A.; MONTORSI, C.; PARRA, B.; POJ, L.; REZZANO, S.; WILLIAMS, L. (2002). *Modelización retórico-lingüística del artículo de investigación científica en inglés e implementación informática de un sistema de redacción asistida*. Informe Proyecto RedACTe II, 06/G179, SeCyT, 1999-2001. Mendoza: UNCuyo.

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.

COPESTAKE, A. (2002). *Implementing Typed Feature Structure Grammar*. CA: CLSI Publications, Stanford University.

DA CUNHA, I. (2004): "Importancia del marcaje de las relaciones discursivas para la generación automática de resúmenes". Actas del VI Congreso de Lingüística General. Galicia: Universidad de Santiago de Compostela.

DIJK, T. A. VAN (1989). *La ciencia del texto*. Barcelona: Paidós.

DUNNING, T. (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19. 61-74.

EDMUNDSON, H. P. (1969). "New Methods in Automatic Extraction". *Journal of the Association for Computing Machinery* 16. 264-285.

FUENTES, M.; GONZÁLEZ, E.; RODRÍGUEZ, H. (2004). "Resumidor de noticies en català del projecte Hermes". Actas del II Congrés d'Enginyeria en Llengua Catalana (CELC'04). Andorra: CELC.

FUENTES, M.; MASSOT, M.; RODRÍGUEZ, H.; ALONSO, L. (2003). "Mixed Approach to Headline Extraction for DUC 2003". Actas del HLT-NAACL Text

Summarization Workshop and Document Understanding Conference (DUC2003). Edmonton, AB, Canada.

FUENTES, M.; RODRÍGUEZ, H. (2002). "Using cohesive properties of text for Automatic Summarization". Actas de las Primeras Jornadas de Tratamiento y Recuperación de Información (JOTRI2002), Valencia, España.

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

GOLDSTEIN, J.; CARBONELL, J.; KANTROWITZ, M.; MITTAL, V. (1999). "Summarizing text documents: sentence selection and evaluation metrics". Actas de 22nd annual international ACM SIGIR conference on Research and development in information retrieval. Berkeley: ACM. 121-128.

HALLIDAY, M.A.K. (1967a). "Notes on Transitivity and Theme in English. Part 1". *Journal of Linguistics* 3. 37-81.

HALLIDAY, M.A.K. (1967b). "Notes on Transitivity and Theme in English. Part 2". *Journal of Linguistics* 3. 199-244.

HALLIDAY, M.A.K. (1967c). "Notes on Transitivity and Theme in English. Part 3". *Journal of Linguistics* 3. 179-215.

HALLIDAY, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold.

HAJIČOVÁ, E.; SGALL, P. (1985). "Towards an automatic identification of topic and focus". Actas de la Second Conference on the European Chapter of the Association for Computational Linguistics. Geneva, Switzerland: ACL. 263-267.

HAJIČOVÁ, E.; SKOUMALOVÁ, H.; SGALL, P. (1995). "An automatic procedure for Topic-Focus identification". *Computational Linguistics*, 21. 81-94,

HAYS, D. G. (1960). *Basic Principles and Technical Variations in Sentence Structure Determination.* Santa Monica, CA: RAND Corporation (Mathematical Division. P-1934).

HAYS, D. G. (1964). "Dependence Theory: A Formalism and Some Observations". *Language*, 40. 511-525.

HLAVACKA, L. (1999). "Diferencias interdisciplinarias en las propiedades lingüísticas de la sección *Method* del artículo de investigación científica en lengua inglesa". *Revista Argentina de Lingüística*. 11-15.

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

HOBBS, J. R.; AGAR, M. H. (1985). *The Coherence of incoherent discourse.* Nota técnica CSLI-85-38, Center for the Study of Language and Information. Stanford, Stanford University.

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

JING, H.; MCKEOWN, K. R. (2000). "Cut and paste based summarization". Actas de la Sixth Applied Natural Language Conference (ANLP-00) and the First Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, Washington: MIT. 178-185.

KITTREDGE, R. (2002). "Paraphrasing for condensation in journal abstracting". *Journal of Biomedical Informatics* 35 (4). 265-77.

KNIGHT, K.; D. MARCU (2000). "Statistics-based summarization – Step one: Sentence compression". Actas de la 17th National Conference of the American Association for Artificial Intelligence. Texas: MIT. 703-710.

KUNO, S. (1972). "*Functional Sentence Perspective: A Case Study from Japanese and English*". *Linguistic Inquiry* 33. 269-320.

KUPIEC, J.; PEDERSEN, J. O.; CHEN, F. (1995). "A trainable document summarizer". Actas de la 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95). New York: ACM. 68-73.

LAMBRECHT, K. (1994). *Information Structure and Sentence Form. Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge: Cambridge University Press.

LENCI, A.; MONTEMAGNI, S., PIRRELLI, V. (2003). "Chunk-it. An Italian shallow parser for robust syntactic annotation"**.** En A. Zampolli, N. Calzolari, L. Cignoni, (eds.). Computational Linguistics in Pisa. *Linguistica Computazionale*. Special Issue, XVI-XVII. Pisa-Roma: IEPI. 353-386.

LIN, C.; HOVY E. (1997). "Identifying Topics by Position". Actas de la ACL Applied Natural Language Processing Conference. Washington: ACL. 283-290.

LÓPEZ ARROYO, B. (2002). "La importancia de la estructuración externa de los *abstracts* en la enseñanza de los lenguajes con fines específicos". Actas de La enseñanza de lenguas en una Europa multicultural, Congreso Internacional de AESLA. Lugo: AESLA. 63-72.

LORÉS, R. (2002). "On the rhetorical structure(s) of abstracts". Actas de La enseñanza de lenguas en una Europa multicultural, Congreso Internacional de AESLA. Lugo: AESLA. 73-80.

LUHN, H. P. (1959). "The Automatic Creation of Literature Abstracts". *IBM Journal of Research and Development* 2. Nueva York: IBM Journal. 159-165.

MANI, I.; MAYBURY, M. (1999). *Advances in Automatic Text Summarization*. The MIT Press.

MANI, I.; GATES, B.; BLOEDORN, E. (1999). “Improving summaries by revising them”. Actas del 37th Annual Meeting of the Association Computational Linguistics. Maryland: ACL. 558-565.

MANI, I. (2001). *Automatic Summarization*. Amsterdam: John Benjamins.

MANN, W. C.; THOMPSON, S. A. (1988). “Rhetorical structure theory: Toward a functional theory of text organization”. *Text* 8 (3). 243-281.

MARCU, D. (1996). “Building up rhetorical structure trees”. Actas de la Thirteenth National Conference on Artificial Intelligence. Oregon: AAAI. 1069-1074.

MARCU, D. (1997a). “From discourse structures to text summaries”. Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 82-88.

MARCU, D. (1997b). “The Rhetorical Parsing of Unrestricted Natural Language Texts”. Actas del 35th Annual Meeting of the Association Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Madrid: ACL. 96-103.

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

MARCU, D. (1999). *Instructions for manually annotating the discourse structures of texts*. Unpublished manuscript. USC/ISI.

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

MEL’CUK, I. (2003). "Levels of Dependency in Linguistic Description: Concepts and Problems". En AGEL, V.; EICHINNGER, L.;. EROMS, H.; HELLWIG, P.; HERRINGER, H. J.; LOBIN, H. (eds). *Dependency and Valency. An International*

Handbook of Contemporary Research. Vol. 1. Berlin - New York: W. de Gruyter. 188-229.

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

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

MIRET, A. M. (2002). "La estructura genérica de la sección "Discussion" en artículos de investigación científica en inglés: un estudio piloto". *Nueva Revista de Lenguas Extranjeras* 7.

MITTAL, V.; BERGER, A. (2000). "Query-relevant summarization using FAQs". Actas del 38th Annual Meeting of the Association for Computational Linguistics (ACL). Hong Kong: ACL.

ONO, K.; SUMITA, K.; MIIKE, S. (1994). "Abstract generation based on rhetorical structure extraction". Actas de la International Conference on Computational Linguistics. Japón: ACL. 344-348.

PAICE, C. D. (1990). "Constructing literature abstracts by computer: Techniques and prospects". *Information Processing and Management* 26. 171-186.

POLLARD, C.; SAG, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: CSLI Publications, University of Chicago Press.

POLLOCK, J.; ZAMORA, A. (1975). "Automatic abstracting research at the chemical abstracts service". *Journal of Chemical Information and Computer Sciences* 15 (4). 226-232.

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

PRADA, J. J. (2001): *Marcadores del discurso en español. Análisis y representación.* Uruguay: InCo, Facultad de Ingeniería, Universidad de la República.

PARDO, T.; NUNES, M.; RINO, M.; (2004): "DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese". Actas del XVII Brazilian Symposium on Artificial Intelligence - SBIA2004. São Luís, Brazil: SBIA. 224-234.

DRAGOMIR R. (1999). *Language Reuse and Regeneration: Generating Natural Language Summaries from Multiple On-Line Sources*. PhD thesis. New York: Department of Computer Science, Columbia University.

DRAGOMIR R.; BLAIR-GOLDENSOHN, S.; ZHANG, Z.; RAGHAVAN R. S. (2001a). "Interactive, Domain-Independent Identification and Summarization of Topically Related News Article". En actas de la 5th European Conference on Research and Advanced Technology for Digital Libraries. London, UK: Springer-Verlag. 225-238.

DRAGOMIR R.; BLAIR-GOLDENSOHN, S.; ZHANG, Z.; RAGHAVAN R. S. (2001b). "NewsInEssence: A System for Domain-Independent, Real-Time News Clustering and MultiDocument Summarization". En actas de la Human Language Technology Conference. San Diego, CA: ACL. 1-4.

ROBINSON, J. (1970). "A Dependency Based Transformational Grammar". Actas del Xme Congrès international des linguistes (Bucarest, 1967) 2. 807-813.

SALAGER-MEYER , F. (1970). "Medical English Abstracts: How Well Are They Structure?". *Journal of the American Society of Science* 42 (7). 528-531.

SGALL, P.; HAJIČOVÁ, E.; PANEVOVÁ, J. (1973). *Topic, Focus, and Generative Semantics*. Kromber: Scriptor.

SGALL, P.; HAJIČOVÁ, E.; PANEVOVÁ, J. (1986). *The Meanings of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel Publishing Company.

SPARCK-JONES, K. (2001). "Factorial summary evaluation". Actas del Workshop on Text Summarization del ACM SIGIR Conference 2001. New Orleans, Louisiana: ACM.

SWALES, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

TESNIERE, L. (1959). *Éléments de syntaxe structurale*. París: Klincksieck [Segunda edición, revisada y corregida, 1969].

TEUFEL, S.; MOENS, M. (1997). "Sentence extraction as a classification task". Actas del ACL/EACL Workshop on Intelligent Scalable Text Summarization. Madrid: ACL. 58-65.

TEUFEL, S.; MOENS, M. (1999). "Discourse-level argumentation in scientific articles: human and automatic annotation". Actas del ACL Workshop: Towards Standards and Tools for Discourse Tagging. Maryland, USA: ACL. 84-93.

TEUFEL, S.; M. MOENS (2002). "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status". *Computational Linguistics* 28. 409-445.

VALLDUVI, E. (1990). *The information component*. Tesis doctoral. University of Pennsylvania.

VON DER GABELENTZ, G. (1869). "Ideen zu einer vergleichenden Syntax: Wort- und Satzstellung". *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* 8. 129-165, 300-338.

WEIL, H (1844). *De l'ordre des mots dans les langues anciennes comparées aux langues modernes.* Boston: Frank.

YZAGUIRRE, Ll.; MATAMALA, A.; CABRÉ, T. (2001a). "El lematizador "PALIC" del IULA (UPF)". *Trabajos en Lingüística Aplicada*. Barcelona: AESLA. 481-485.

YZAGUIRRE, Ll.; MATAMALA, A.; BACH, C.; CASTILLO, N.; USTRELL, E. (2001b). "AMBILIC, el desambiguador lingüístico del Corpus del IULA (UPF)". *Trabajos en Lingüística Aplicada*. Barcelona: AESLA. 473-480.