Universidad de Granada
Facultad de Comunicación
y Documentación

Universidad de la Habana
Facultad de Comunicación

Título:

# Texminer: Un modelo para la extracción y desambiguación de textos científicos en el dominio de Ingeniería de Puertos y Costas

## Tesis Doctoral

Autor:
MSc. Amed Abel Leiva Mederos

Director:
Dr. José Antonio Senso Ruiz

Granada, Julio 2011

Editor: Editorial de la Universidad de Granada
Autor: Amed Abel Leiva Mederos
D.L.: GR 1412-2012
ISBN: 978-84-695-1055-1


these things
types outside himself
called τὸ ἀρχέτυπον φῶς
times in Dionysius
hierarchia, II, 4
nominibus, I,
is not found
diversis quaestio
which are
understa
Plato

# Exergo

# Exergo

*La complejidad nos prepara para vivir lo inesperado, aunque no nos libra de la incertidumbre.*

*Edgar Morín*


# Dedicatoria

# Dedicatoria

*A mi padre por qué esté donde esté yo seré siempre su mayor orgullo.*

*A mi madre: Dra. Inés María Mederos Morell por inculcarme confianza y valor.*

*A tía Dulce y abuela Micalea por todo el amor que me dieron siempre.*


# Agradecimientos

# Agradecimientos

Un trabajo de esta envergadura lleva el conc urso de muchos profesionales, es por ello que quisiera agradecer a las siguientes personas:

Al Dr. José Antonio Senso Ruíz, por guiar mis pasos en el terreno de investigación, por dedicar todo su esfuerzo para que aprendiera las tecnologías de la Web Semántica y por sus atinadas sugerencias en la presentación del informe, además por ser mí amigo.

A la Dra. Ania Ro sa Hernández Quintana, por formar parte de este em peño y sugerir el tema de investigación.

Al MSc. Sandor Eduardo Domí nguez Velasco, por el desarrollo del software y por soportar todas mis exigencias durante tres años.

Al MSc. José Rivero Días, por confiar en e sta loca id ea y facilit ar toda la logística para el desarrollo de la investigación.

A mis profesoras: Ester Hernández, M ilagros Velás quez, Magdalen a Mena y Aleida Mena

A la MSc. María Sofía Claro Duyos mi maestra de matemáticas en la Universidad.

Al Lic. Eric k Olivera Acos ta, al Diseñador Sergio Alej andro Rodríguez, y al MSc. Michel Aliosha Pérez Gonzáles, con un equipo de trabajo como el vuestro todo puede lograrse.

A la MSc. Diana María Fernández Moreno, la persona que me ense ñó a hacer bases de datos.

Al Proyecto Puertoterm de la Universi dad de Granada, por colaborar con esta investigación.

A los colegas del Departamento de Ciencia s de la Información y de la Facu ltad de Ciencias de la Información, en especial a Damaris, Sara y Belkys Rodríguez, ustedes me han visto crecer en toda mi vida profesional.

A Julia Ross Rojas, por ayudarme tant o en mi vida profesi onal y facilitarme siempre buenos consejos.

A la MSc . Vadia Estévez Chaviano, por sus revisiones de los aspectos metodológicos de la tesis.

# Agradecimientos

Al Grupo de Servicios especializados del CDICT de la Universidad a Marilyn, Ohilda, Vladimir, Esperanza y Zoe.

A mis com pañeros del Máster, Manolo, Maria Josefa Peralta, Floriselda, Anisley, Ansleiby, Gladis, Keilyn y Yudeisy y Darlin.

# Resumen

El Procesamiento del Lenguaje Natural se ha convertido desde los años 40 en una herramienta para el desarrollo y tratami ento de resúmenes automáticos. Múltiples disciplinas se han encar gado de la formulación y la implementación de modelos para la extracción de texto, entre ellas se encuentr an: la Cibernética, la Lingüística, la Semiótica, la Cibersemiótica y la Ciencia de la Información. En esta investigación se propone un Modelo para la Extracción y Desambiguación d e Textos Científicos en el dominio de la Ingen iería de Puertos y Costas para facilitar la Representación de la Información textua l y la Recuperación de la Información en el referido campo. Para desarrollar el modelo se realiza un abordaje teórico, conceptual y metodológic o que engloba el terreno del resumen automático. A partir del análisis teórico y del concurso de métodos de investigación emanados de la Psic ología Cogn itiva, la Te rminología y la Ciencia de la Informaci ón se construye el modelo propuest o en la in vestigación. Para s u validación se construye una herramienta de software, sustentada sobre r eglas disc ursivas emanadas del análisis lingüísti co de 50 text os asociados al dominio objeto de la investigación. Se ev alúa el modelo te niendo en cuent a diversos puntos de vista entre los que se dest acan: usabilidad de la herramienta, cali dad de la ontología, calidad de los resúmenes obtenidos y la calidad de los corpus. Finalm ente se arriba a conclusiones, recomendaciones y se exponen aquellas aportaciones que se han logrado con la investigación de este tema.

# Índice

# Índice


# Introducción

# Introducción

A diferencia de la indización y la clasificación, los resúmenes son una reelaboración textual[1], una propuesta por reducción significativa de los originales que parte de algunas cuestiones previas, como las características del texto a resumir, el destinatario del resumen, los objetivos que se plantee y el proceso que le da lugar, y datan, como servicio organizado, desde mediados del siglo XIX Hernández (Hernández, 2007).

Los resúmenes se erigen como metadatos, pues son la expresión que enuncia el contenido de los documentos textuales Herrera (Herrera, 2007). En la literatura de la Ciencia de la Información constantemente se hace referencia a sistemas de metadatos como: Dublin Core, EAD, ISAD-G, MARC, MARC 21, etc. Sin embargo, en pocos casos se habla del resumen documental como metadato que representa de forma resumida el contenido de los textos, pues es poco su reconocimiento como estructura representacional de datos no estructurados. Esta forma de procesamiento de la información ha sido tratada tradicionalmente como un proceso netamente lingüístico, computacional y semántico.

El resumen documental es uno de los elementos, que dentro de la etapa de la Representación de la Información sirve para la mediación textual. Su desarrollo puede constatarse a través del surgimiento de las primeras descripciones formales de contenido en las bibliotecas de la antigüedad. Aunque su evolución desde el punto de vista técnico comenzó en los albores del desarrollo bibliotecario, es cierto que su auge como servicio informativo y herramienta de representación documental viene dado justamente por un fenómeno que condiciona el desarrollo de la Ciencia de la Información: La Explosión Documental ocurrida en la década del 60 del pasado siglo. Los años 90 del siglo XX, traen como novedad la generalización de la red de redes. Este suceso trae nuevos cambios en el área de la comunicación convirtiéndose en un nuevo escenario para la producción científica. El resumen documental adquiere alta connotación con la expansión de Internet y la aparición de los documentos

[1] Se construye otro texto, con todas sus implicaciones.

digitales. Es entonces que surge la necesidad de resumir la información que circula en la red, la cual tiene diversas características.

El desarrollo de Internet trae consigo el enfrentamiento a nuevas problemáticas en la representación de contenidos. Los nuevos documentos que circulan en las estructuras reticulares presentan plataformas lógicas que se actualizan mediante mecanismos de inteligencia artificial y además combinan en su estructura texto, sonido e imágenes, que se armonizan en un nuevo entorno de comunicación con diversas características semiológicas.

El creciente protagonismo de los documentos hipertextuales plantea nuevas propuestas documentales que difieren de las tradicionales estructuras textuales. Con las nuevas formas de texto, la noción de documento se sale de la línea tradicional. El marco referencial y de reenvíos no vincula sólo documentos, sino también otras estructuras mucho más complejas que poseen características multimediáticas.

La producción automática de resúmenes, sin embargo, no ha logrado superar las estrategias inferenciales de los resumidores humanos, a pesar de ser la base para una red semántica extensiva que en la actualidad se materializa como autosumarios o autoextractos de oraciones clave.[2] Los resúmenes se han transformado en un metadato en la Red, siguiendo los modelos estructurados de generación de documentos, o en una "plantilla" de estructuras retóricas, sin embargo, aunque ciertos motores especializados incorporen resúmenes de esta manera, esto no ocurre así para todos los recursos electrónicos involucrados.

Para Pinto (Pinto, 2001) los clásicos sistemas documentales estructurados en forma (unidireccional/lineal y jerárquica) evolucionan a formas más complejas bidireccionales/reticulares/democráticas).

La aparición de la Web y la documentación digital agilizó y transformó los servicios y los procedimientos de construcción de resúmenes a escala global. Sin embargo, aún los resultados en materia de métodos de extracción de textos

[2] Estos sumarios y extractos son microtextos representativos que se generan de forma automática y se insertan en la praxis hipertextual.

no son los esperados, pues subsisten dificultades con el balance textual, la cohesión y la coherencia. La Ciencia de la Información tiene ante sí el reto estudiar (de conjunto con la Cibernética, la Lingüística, la Semiótica, la Psicología, etc.) estos problemas, pero la realidad indica que son muy pocos los estudios teóricos que desde la Ciencia de la Información abordan esta temática si se compara la producción con la Ciencia de la Computación.

El resumen como modelo de representación ha sido el más empírico de los procesos de Representación de Información, y a la vez el menos conceptuado desde las posturas disciplinares. Abundan definiciones asociadas a su brevedad, tipología y funcionalidad, pero su estudio más completo de cara a las nuevas expectativas en el entorno digital casi con seguridad ocurre en Pinto (Pinto, 2001), en el que se destaca su textualidad, la importancia del dominio discursivo y operatividad contextual.

El principal cambio que ha experimentado el resumen desde su surgimiento se encuentra en el uso, pues los resúmenes han pasado del entorno secundario al primario, insertándose en los recursos como microtextos que facilitan el hojeo en lugar de la búsqueda tradicional, transformándose en un tipo de estructura con carácter visualizador-navegador. De este modo, la estructura de un párrafo tradicional evoluciona a una modular, en múltiples párrafos acompañados por subencabezamientos. Sin embargo, servicios de resúmenes especializados como el *Chemical Abstracts* (*CAS Online*) continúan sirviéndose de metodologías clásicas de extracción textual.

La discusión sobre las perspectivas del resumen en la ciberrealidad apenas está comenzando, lo que implica para estos modelos la traslación de un enfoque centrado “en la lingüística original hacia otras aproximaciones de índole pragmática”, que ven al resumen como “un sistema estratégico de producción textual que ya no puede deducirse unívocamente desde el texto fuente sino que depende de determinadas expectativas y necesidades de los usuarios.” Esto hace que más importante que la normalización de los resúmenes sea la necesidad de fomentar su “equivalencia funcional”, la capacidad de adaptar cada

resumen a la situación contextual desde una visión estratégica Pinto (Pinto, 2001).

La producción automática de resúmenes se ha basado en métodos estadísticos desde las investigaciones de Lunh (Lunh, 1958) y continúa hoy con los trabajos del Procesamiento del Lenguaje Natural, las aplicaciones basadas en analizadores automáticos y en modelos socio-pragmáticos del discurso, basados en técnicas de "dominio de personalidad propia", subyugados por la extracción de palabras o cadenas significativas, que ya en los años 90 recortan las "estructuras retóricas" Endres Niggemeyer, Mann y Thompson (Endres-Niggemeyer, 2005, Mann and Thompson, 1990, Endres-Niggemeyer, 1995), muy relacionadas con el contenido de los documentos, al aportar indicios sobre el esquema conceptual y organizador de las distintas unidades que conforman un texto para reproducir y ordenar determinadas frases Ono (Ono et al., 1994).

Las últimas tendencias en el terreno de la lingüística computacional son evidencia del poco desarrollo en la creación de modelos que logren identificar al usuario, sus necesidades y su realidad comunicativa. Solo algunos sistemas combinan ontologías, tesauros y herramientas de recuperación de la información de alto nivel de sofisticación como los llamados modelos de desambiguación, sin embargo los resultados siguen siendo los mismos que en la década de los 90, debido a que la producción industrial de herramientas para los procesos sumistas sigue siendo émula de teorías donde se excluye el análisis de dominios lingüísticos, con los que se logra un mayor balance en el resultado de la extracción del texto.

Los modelos extractivos en que se sustentan hoy las herramientas para la confección de resúmenes automáticos son monodimensionales, en su mayoría, pues su consistencia teórica solo observa las posturas lingüísticas, matemáticas y tecnológicas, ignorando las teorías comunicativas y la aparición de nuevas formas de texto en los medios electrónicos. Esto hace que se necesite un modelo de resumen desarrollado desde la BCI (Bibliotecología y la Ciencia de la Información).

## Justificación

La necesidad de desarrollar desde la Ciencia de la Información y la Bibliotecología modelos para el desarrollo de la extracción de textos se debe a que el uso de modelizaciones y paradigmas clásicos para el tratamiento del texto verbalizado no están acordes a las exigencias de los procesos actuales de Representación de Información y Conocimiento, por tanto para desarrollar nuevos enfoques en esta área se necesita de noveles modelos para interoperar los sistemas de Representación Textual. Se estima que son insuficientes e ineficientes los modelos que existen para el desarrollo de la extracción de texto debido al apego que mantienen a reglas clásicas, basadas en modelos matemáticos y lingüistas amparados por procederes y visiones reduccionistas del campo de la Representación Textual. Una Representación Textual debe estar sustentada en diversas visiones que involucren varias disciplinas como: la semiótica, la lingüística, los modelos cognitivos, la matemática y la computación. Por ello se impone una nueva modelización que permita desde la Ciencia de la Información y la Bibliotecología formular un modelo que propicie la extracción formalizada de textos que circulan en la red de redes. Esta situación da pie al problema de investigación que se explica a continuación. Por otra parte numerosos investigadores de jóvenes disciplinas como la: Ingeniería de Puertos y Costas carecen de herramientas para la extracción de información, lo que contribuiría al desarrollo de las potencialidades investigativas en este campo del conocimiento.

## Problema de Investigación

### Situación Problémica

No existe en el terreno de la Ingeniería de Puertos y Costas un modelo de resumen que facilite la extracción y desambiguación de textos científicos, lo cual causa que los profesionales que se dedican a este campo carezcan de servicios de información que les permitan obtener información resumida de la red de redes. De esta situación Problémica se desprende el siguiente problema de investigación:

¿Qué elementos son necesarios considerar para la confección de un modelo socio-cognitivo para la extracción y desambiguación de textos científicos en el dominio Ingeniería de Puertos y Costas?

***Objetivos***

Los objetivos de la investigación están consonancia con la complejidad de la investigación y los resultados que de ella se esperan.

**Objetivo General**

- Proponer un modelo para la confección de resúmenes automáticos acorde a las exigencias de la Ciencia de la Información en el dominio Ingeniería de Puertos y Costas.

**Objetivos Específicos**

- Estudiar el entorno teórico que ha afectado la confección de resúmenes automáticos.
- Estudiar las normas y procedimientos que propician la construcción de resúmenes automáticos.
- Valorar los software y los métodos para el desarrollo de resúmenes por vía automática.
- Diseñar un Modelo para la elaboración automática de resúmenes en el do minio Ingeniería de Puertos y Costas.
- Estudiar el discurso de la Ingeniería de Puertos y Costas, estableciendo particularidades de su estructura textual de modo que las regularidades de redacción y construcción se conviertan en heurísticas.
- Implementar el modelo para la extracción y desambiguación de textos científicos en el dominio de Ingeniería de Puertos y Costas.
- Evaluar el modelo propuesto a través del sistema de extracción y desambiguación de textos científicos aplicado al dominio Ingeniería de Puertos y Costas.

**Interrogantes Científicas**

¿Cuáles son los referentes teórico metodológicos que sustentan la confección de resúmenes científicos?

¿Qué procedimientos y normas son los más efectivos para la confección de resúmenes automáticos?

¿Qué cualidades de inferencia tienen los software para la confección automática de resúmenes?

¿Qué parámetros son necesarios para la creación de un método para la construcción de resúmenes en el terreno de la Ingeniería de Puertos y Costas**?**

¿Qué modelo puede elaborarse para la confección de resúmenes en un espacio virtual acorde a las necesidades de la Ingeniería de Puertos y Costas y dentro de la demandas de las Ciencias de la Información?

¿Qué características distinguen al discurso de la Ingeniería de Puertos y Costas?

¿Qué características debe tener una aplicación desarrollada a partir del modelo propuesto para lograr la extracción y desambiguación de textos en el dominio de Ingeniería de Puertos y Costas?

¿Cuáles son los resultados de la aplicación del modelo de extracción y desambiguación en el terreno de la ingeniería de Puertos y Costas?

**Estructura Capitular**

La tesis está descrita en 10 Capítulos que se explican a continuación:

**Capítulo I:** Marco Metodológico de la Investigación: Se describen y fundamentan los métodos de investigación desarrollados en la investigación. Los epígrafes que incluye este capítulo son los siguientes:

- ***Epígrafe 1 Contexto de la Investigación,*** es un segmento de la investigación dedicado a declarar, que el estudio se realiza en Cuba, en la Universidad Central de las Villas, utilizando la terminología del proyecto Puertoterm de la Universidad de Granada, y que los participantes en los

procesos investigativos son especialistas tanto cubanos como internacionales.

- ***Epígrafe 2, Variables, operacionalización,*** describe las variables utilizadas en la investigación y sus dimensiones operacionales
- ***Epígrafe 3: Objeto de Investigación,*** dedicado a fundamentar el objeto de la investigación desde la dimensión de las Ciencias de la Información.
- ***Epígrafe 4: Perspectiva de la Investigación***, donde se declara y se justifica el corte mixto de este tipo de estudio, mediado por el uso de técnicas cuantitativas y cualitativas.
- ***Epígrafe 5: Métodos de la investigación:*** Una exposición justificada de todos los métodos de investigación utilizados en la investigación, además de la explicación de la forma en que se realizan todas la técnicas y los procedimientos.
- ***Epígrafe 6: Etapas de la Investigación,*** se refiere a todas las etapas de la investigación.
- ***Epígrafe 7: Herramientas de Programación y Modelación:*** Se centra en la exposición de todas las herramientas de programación, de marcación y de edición de utilizadas en la tesis
- ***Epígrafe 8: Análisis de las Fuentes de Información:*** Descripción pormenorizada de las fuentes de información, utilizadas en la tesis, estudio de los autores principales en el tema en el tema de resumen automático.

**Capítulo II.** El Resumen automático: fundamentos teóricos y metodológicos para su construcción. En este segmento de la investigación se aborda sobre el desarrollo del resumen automático, su definición conceptual, métodos de evaluación, los retos del resumen en la sociedad digital, la normalización, el uso de agentes automáticos, paradigmas dominantes en el resumen y sus metodologías. Los epígrafes de este capítulo son los siguientes:

- ***Epígrafe 1: Definición del concepto de resumen:*** Se aborda el concepto de resumen documental a partir de la posición de varios autores, demostrando el desarrollo y evolución de este concepto a la luz de las nuevas tipologías textuales que aparecen en la red de redes (INTERNET)
- ***Epígrafe 2: La Sociedad Digital:*** Refiere las connotaciones de la sociedad digital , sus características y la viabilidad del uso de los resúmenes en esta sociedad
- ***Epígrafe 3: Normalización y Calidad en el Resumen automático: métodos de Evaluación:*** Se centra en el estudio de las normas para la construcción de resúmenes a través de un análisis crítico de los modelos normalizativos. Este análisis conlleva a determinar que la normalización estructural no es una postura viable en el resumen para el entorno de la Web.
- ***Epígrafe 4: Los algoritmos matemáticos : su función en la evaluación de Corpus y en la Selección de Términos***: Una panorámica de los algoritmos de agrupamiento de texto traídos de la Minería de Datos y los algoritmos específicos para el desarrollo de resúmenes se trata en este acápite, donde se valoran los mejores algoritmos que han surgido después de las implementaciones de Salton hasta los trabajos de Inderjeet Mani utilizados generalmente en la evaluación de corpus y en cualificación de términos.
- ***Epígrafe 5: Los agentes automáticos en el procesamiento del lenguaje natural:*** Dedicado por completo al uso de agentes en el procesamiento de información textual como entes reproductores de estrategias cognitivas.
- ***Epígrafe 6: Paradigmas del resumen automático***: Una visión crítica de los paradigmas dominantes del resumen automático a través de sus tres posturas: cognitivo, físico y socicognitivo.

- ***Epígrafe 7: El resumen como proceso y producto documental:*** Detalla los procesos que hacen que el resumen automático esté en la doble función proceso-servicio.
- ***Epígrafe 8: Procesos que intervienen en el trabajo con el resumen automático***: Describe cada uno de los procesos que se integran al resumen automático a partir de una valoración sistémica y holística con una concepción teórica que exige la Ciencia de la Información.
- ***Epígrafe 9: Desarrollo de las técnicas de extracción de texto verbalizado.*** Un recuento de las técnicas de extracción de de texto verbalizado teniendo en cuenta la existencia de procesos de Organización y Representación del Conocimiento, entre los que se encuentran: Topic Maps, Ontologías y Mapas Conceptuales y Mapas Mentales.
- ***Epígrafe 10: Metodología del resumen automático:*** Se declara la metodología seguida en el desarrollo del resumen automático y se insiste en la asunción de nuevos modelos.
- ***Epígrafe 11: Retos de los sistemas de representación textual ante el resumen automático***: Una valoración crítica de los retos y los cambios que deben implementar los sistemas de Representación de la Información y el Conocimiento ante los nuevos paradigmas que impone el resumen en el entorno digital.
- ***Epígrafe 12: El resumen automático: retos investigativos*.** Reflexión de los derroteros que tiene ante sí la investigación sobre los aspectos metodológicos del resumen a partir de una óptica multidisciplinar.

Capítulo III. ***Los métodos de construcción automática de Resúmenes de Texto***, dedicado exclusivamente al análisis de los métodos y las herramientas de construcción de textos, haciendo énfasis en los algoritmos de agrupamiento y analizando las distintas formas e implementación de los resúmenes textuales desde una perspectiva operativa.

- ***Epígrafe 1: Componentes de Un Sistema de Resumen:*** Proposición de los componentes de un Sistema de Resumen.
- ***Epígrafe 2: La organización del Texto en los contextos digitales:*** Un análisis de las formas de organizar el texto en los medios digitales y sus influencias en los modelos de tratamiento de texto.
- ***Epígrafe 3: Análisis de Dominio y Extracción de Texto:*** Se centra en la preponderancia del Análisis de Discurso y los estudios de dominio como clave de organización y mecanismo de construcción heurística para el desarrollo de los resúmenes. Se asume el concepto de estudios de necesidades a partir de la visión lógica de los estudios de usuarios y en la implementación de servicios y productos.
- ***Epígrafe 4: Procesos fundamentales en la producción de Resúmenes:*** Explicación de los procesos fundamentales para la producción de resúmenes automáticos.
- ***Epígrafe 5: Métodos de extracción de Texto Verbalizado:*** Presenta un análisis de las técnicas de confección de extractos, desde los métodos de índole estadística hasta los métodos híbridos.
- ***Epígrafe 6: Técnicas descritas para la confección de Resúmenes de Texto:*** Estudio de las estrategias para implementar resúmenes de texto, señalando las tipologías de estos documentos cuando su naturaleza es automática.
- ***Epígrafe 7: Complejidades del tratamiento del Resumen Multidocumento:*** Una valoración del tratamiento del resumen de documentos múltiples a partir de una visión cibernética de sus posibilidades de automatización.
- ***Epígrafe 8:*** Explicitación de las técnicas para desambiguar léxicos y sus posibles aplicaciones en la Minería Textual.

- ***Epígrafe 9: Software para la construcción de resúmenes de Texto:*** Identificación de aquellos software y sistemas que desarrollan servicios de Minería Textual.
- ***Epígrafe 9: El Tratamiento del Texto en Cuba***: Revisión del estado de la investigación sobre tratamiento e implementación automática de resúmenes en Cuba a partir de una revisión crítica.
- ***Epígrafe 10: Consideraciones sobre los métodos de Resumen Automático:*** valoración de los métodos de resumen automático y sus retos en la Representación de la Información y el Conocimiento.
- ***Epígrafe 11: Aportes de la Ciencia de la Información para desarrollo de los Sistemas de Resúmenes***: Aportes de la Ciencia de la Información a la concepción de sistemas de resúmenes de texto mediante la visión de especialistas de diversas ramas del saber.

**Capítulo IV. TEXMINER:** Metamodelo para la extracción y desambiguación de textos científicos, se encarga de mostrar el modelo de resumen propuesto y todas sus etapas, señalando ejemplos de implementación, así como las premisas teóricas y los componentes metamodélicos.

- ***Epígrafe 1: Modelo conceptual propuesto que permite el resumen automático de textos:*** Declara el sustento teórico del modelo, así como sus principios esenciales.
- ***Epígrafe 2: Modelo Empírico aplicado a la UCLV:*** Presenta toda la investigación cognitiva previa al modelo, específicamente la detección de estrategias cognitivas en resumidores de 12 facultades de la UCLV.
- ***Epígrafe 3: Pasos del Modelo:*** Una descripción de todos los pasos del modelo, desde la entrada de los datos hasta la salida del resumen.
- ***Epígrafe 4: Competencias necesarias para la aplicación del modelo:*** Destaca las competencias que un profesional de la información demanda para enfrentarse a este tipo de sistema de Representación de Información.

**Capítulo V.** Análisis del Discurso en el dominio de Ingeniería de Puertos y Costas, se encarga de mostrar el análisis de los textos del corpus del proyecto Puertoterm[3] de donde se declaran las reglas discursivas que sirven para la desambiguación, extracción y construcción de resúmenes.

- ***Epígrafe 1: Criterios de Selección de Corpus:*** Se edifica la teoría de selección de Corpus y los principios en que se basa el autor para seleccionar el corpus de estudio de la Investigación.
- ***Epígrafe 2: Características del Proyecto Puerto Term.*** Expone las características, los principios y objetivos del proyecto Puertoterm, así como las características del Corpus de la investigación.
- ***Epígrafe 3: Herramientas de Selección de Corpus:*** Analiza y describe todas las herramientas de selección de corpus, reflejando los valores del software de pago Word Smith toolls.
- ***Epígrafe 4: Análisis de Discurso en el dominio de la Ingeniería de Puertos y Costas:*** Se proponen los indicadores para la realización del análisis de discurso en el terreno de la Ingeniería de Puertos y Costas, a partir de las posibilidades que brinda el estilo, la semántica y la retórica textual.
- ***Epígrafe 5: Análisis de Corpus:*** Dirigido exclusivamente al estudio del Corpus, en su composición macro estructural y microestructural, estudiando sintagmas, palabras, posiciones de vocablos y partes del párrafo.
- ***Epígrafe 6: Formalización de Reglas para la Extracción del Texto:*** Expone todas las reglas de discurso declaradas a partir del estudio del corpus.
- ***Epígrafe 7: Criterios de desambiguación léxica:*** Incluye los procesos esenciales que se desarrollan para desambiguar las palabras y darles sentido pragmático una vez que está listo el texto en su fase extractiva.

[3] Proyecto de Ingeniería de Puertos y Costas en que se sustenta la investigación

- ***Epígrafe 8: Retos de Implementación de las Reglas Textuales.*** Se dedica a reflexionar sobre los retos que demanda la implementación de reglas heurísticas desarrolladas a partir del discurso.

**Capítulo VI.** ***Puertotex***: un sistema para la extracción y desambiguación de textos científicos en el dominio de la Ingeniería de Puertos y Costas. En este capítulo se muestra la implementación de las reglas discursivas y por tanto la aplicación del modelo propuesto a partir de una herramienta ad hoc. Se hace insistencia en el diseño de servicios de información, los módulos del sistema, la selección del lenguaje de programación para hacer la herramienta, la construcción de ontologías, base de datos finalmente se presenta todo el sistema documentado mediante técnicas de ingeniería de software y los manuales de usuario y los de instalación del sistema.

- ***Epígrafe 1: Estudio de Necesidades: Resultados:*** Presenta el estudio de necesidades que servirá de eje central al desarrollo el sistema, para ello detalla las variables que se demandan para la confección del sistema.
- ***Epígrafe 2: Implementación de Reglas Textuales:*** Expone todos los datos XML que el sistema usa, teniendo en cuenta la DTD, desarrollada a tales efectos. Se especifican en este capítulo los diversos software y lenguajes de programación utilizados para el trabajo con la Web semántica y se detallan las peculiaridades del XML.
- ***Epígrafe 3: Construcción de la Ontología y las Bases de Conocimiento:*** Es un segmento de la investigación dedicado exclusivamente a las formas de construcción de la ontología, donde se muestran todas las clases de la ontología, los axiomas, las anotaciones, las reglas de inferencia, las propiedades de los objetos y de las clases. También se describen en este capítulo las herramientas de anotación y construcción de ontologías.
- ***Epígrafe 4: Agentes:*** Se explican los modelos para el desarrollo de agentes en el sistema y su forma de implementación.

- ***Epígrafe 5: Metodología de implementación del Sistema:*** Se dedica a la explicación de los procesos de Ingeniería de Software.

**Capítulo VII.** ***Análisis de los Resultados de la Implementación del Modelo***, esta parte de la investigación se refiere a un conjunto de procesos de índole matemático, cuantitativo y cualitativo que permiten evaluar la ontología, la calidad de las palabras, los resúmenes resultantes, la usabilidad del sistema y la valoración final del modelo por los expertos. Los epígrafes que incluye esta parte son los siguientes:

- ***Epígrafe 1: Evaluación de Corpus:*** Se aplican técnicas para la validación del Corpus, logrando con esto la selección del corpus para el tratamiento léxico y del corpus patrón para el diseño del sistema.
- ***Epígrafe 2: Evaluación de Ontologías:*** Se propone un modelo para la evaluación de la calidad de la Ontología Ontosatcol a partir de indicadores léxicos, de recuperación de información y de estructura sintáctica.
- ***Epígrafe 3: Evaluación Matemática del Modelo:*** Evalúa el modelo a partir de N-gramas, utilizando para ello la herramienta Rouge. Compara el resumen obtenido con nuestro modelo a con otros modelos.
- ***Epígrafe 4: Evaluación del Sistema por especialistas:*** Se valora el sistema a partir de un test de evaluación heurística desarrollado por especialistas en el tema.
- ***Epígrafe 5: Evaluación del Modelo por criterio de Expertos:*** El modelo se valora a través de sus etapas y presupuestos teóricos a través de expertos.

**Capítulo VIII.** Conclusiones, se declaran las conclusiones de la investigación.

- ***Epígrafe 1: Conclusiones:*** Se exponen la conclusiones de la investigación.

**Capítulo IX**. Trabajo Futuro, este punto de la investigación se encarga de definir las líneas de trabajo futuro que habrán de llevarse con esta investigación.

- ***Epígrafe 1: Trabajo Futuro:*** Se formulan propuestas de trabajo futuro y nuevas recomendaciones para el desarrollo del tema de investigación.

***Capítulo X. Aportes de la Tesis***: Muestra las publicaciones y eventos en los que participó el autor a partir del tema de investigación.

La norma usada para la descripción bibliográfica de las fuentes de información fue el estilo Harvard.


# Referencias Bibliográficas

## Referencias Bibliográficas

ENDRES-NIGGEMEYER, B. 2005. SimSum: an empirically founded simulation of summarizing *Information Processing and Management,* 36**,** 659-682.

ENDRES-NIGGEMEYER, B., MAIRE, E. Y SIGEL, A. 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing & Management,* 31**,** 631-674.

HERNÁNDEZ, A. 2007. *Paradigmas dominantes de la Representación de la Información y el Conocimiento.* Universidad de la Habana.

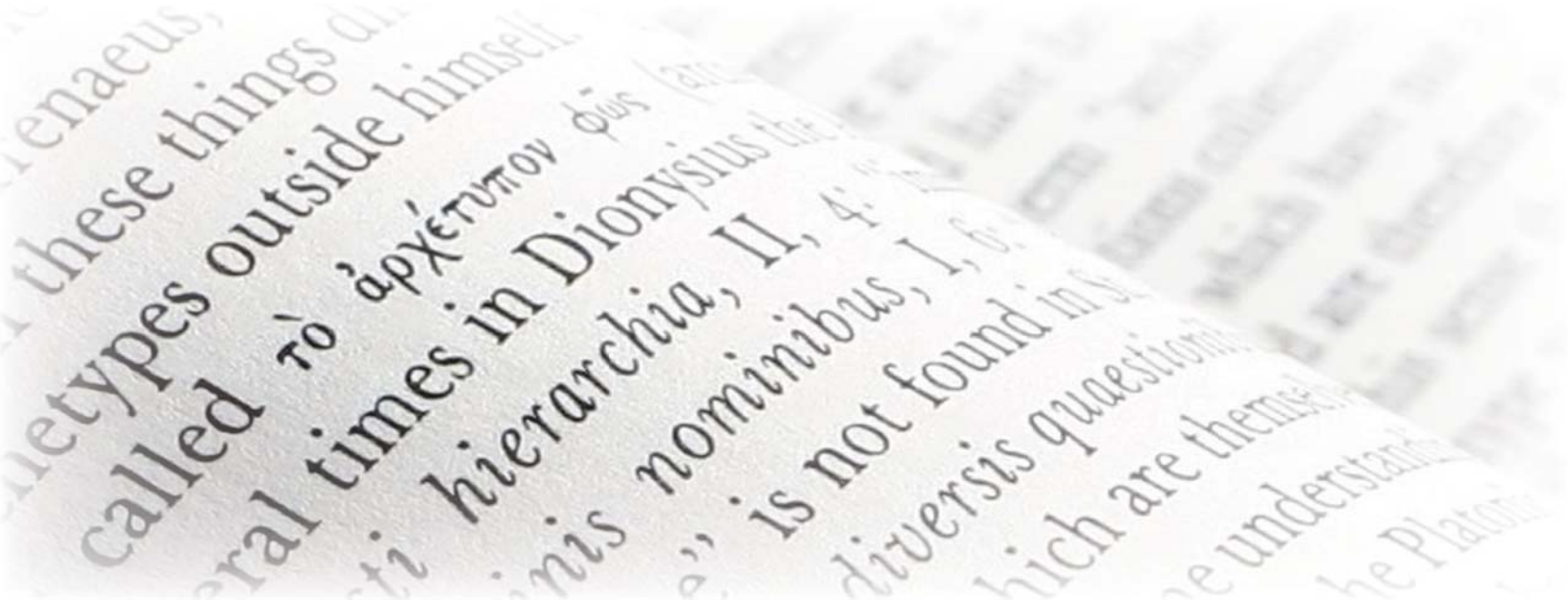
HERRERA, R. 2007. *Formatos de Comunicación* [Online]. La Habana: Universidad de la Habana, Cuba. Available: http://fcom.uh.edu.cu [Accessed 16.julio 2009].

LUNH, H. 1958. The Automatic creation of Literature abstracts. *Journal of Research of Development***,** 159 – 165.

MANN, W. & THOMPSON, S. 1990. Rhetorical structure theory: a theory of text organization.

ONO, K., SUMITA, K. & MIIKE, S. 1994. Abstract generation based on rhetorical structure extraction. *In:* Proceedings of the International Conference on Computational Linguistics, Kyoto.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez




# Capítulo I

## Marco Metodológico de la Investigación

# Capítulo I. Marco Metodológico de la Investigación

## 1.1.- Introducción

La investigación que se presenta ha sido desarrollada utilizando diversas técnicas y procedimientos de índole matemático, cognitivo y lingüístico. Construir desde la academia un modelo para extraer texto, usando como mecanismo de estructuración las características del dominio es algo poco conocido en la literatura de la especialidad. Este capítulo inicial de la tesis doctoral se centra en la explicitación de aspectos medulares para la comprensión del estudio llevado a cabo en esta investigación. Los acápites en que se divide el capítulo son los siguientes:

Operacionalización de las Variables

Objeto de la Investigación

Campo de Investigación

Perspectiva de la Investigación

Métodos de Investigación

Estudio de autores y fuentes de información

Limitaciones

Es importante destacar que en esta investigación se toma como base de análisis los Corpus Textuales del Proyecto Puertoterm de la Universidad de Granada, para la validación e implementación del modelo propuesto.

### 1.1.1.- Contexto de la Investigación

Como toda investigación que se ampara en una perspectiva mixta, esta se centra en un contexto investigativo particular, la Facultad de Ciencias Agropecuarias de la Universidad Central" Marta Abreu" de las Villas, el proyecto PUERTOTERM de la Universidad de Granada, España y Bibliotecarios Especialistas de las Bibliotecas de Facultades de la UCLV.

El proyecto PUERTOTERM fue el que sirvió de base terminológica para el estudio seleccionado, por ser un dominio donde se aplican muchas disciplinas relacionadas con la Ingeniería, las Ciencias Puras, las Ciencias Sociales y la Geografía, mezcla de la Ingeniería Naval y la Ingeniería de Caminos, además de ser un espacio donde se inserta una gran cantidad de profesionales de diversas áreas disciplinares, por ello se tomó como sustento lo que plantea Meneses Placeres (Meneses Placeres, 2010) cuando declara que en una investigación que utiliza también el enfoque cualitativo se exige que el investigador se coloque en la situación que mejor le permita recoger información relevante para el concepto o teoría buscada. Teniendo en cuenta esto, se han planteado las siguientes comunidades de estudio[4]:

- Especialidad de Ciencias biológicas, definidas como dominios que trabajan con información ecológica, biológica, física, etc.
- *Especia*lidad en Ingeniería de Puertos y Costas. Este es un dominio compuesto solo por expertos en Ingeniería Civil, Ingeniería Mecánica y Naval, desarrollado a partir de los documentos que pertenecen a una comunidad de discurso, específica, reconocida por la integración de conocimientos inherentes a la Ingeniería de Puertos, Ingeniería Naval e Ingeniería Civil.
- Lingüistas: Expertos en el estudio del léxico son los que observan las categorías lingüísticas en los corpus de la investigación para su selección.
- Psicología: Comunidad que maneja el estudio de las habilidades cognitivas dentro de la investigación, a esta pertenecen tanto pedagogos como psicólogos de profesión.
- Ciencias de la Información, comunidad de práctica donde se desempeñan profesionales y especialistas que se dedican tanto a la docencia como a las tareas de trabajo intensivo con la información (servicios y tareas del trabajo de las organizaciones informativas). El doctorando pertenece a la

[4] Entiéndase como todo sujeto u objeto que participe en la investigación.

comunidad Ciencias de la Información, en su posición de docente en el departamento de ciencias de la información y en posición como metodólogo en el centro de Información y Documentación de la UCLV.

- Usuarios Generales: Se denomina usuarios generales a otros individuos que no son expertos en el tema, pero pueden usar el sistema que se propone para su actividad investigadora, dentro del universo de la Universidad Central" Marta Abreu" de las Villas son Usuarios Generales los estudiantes y profesores de aquellas carreras que no son propiamente de las temáticas que abarca la Ingeniería de Puertos y Costas.

Se utilizaron diversas muestras y poblaciones completas en la investigación cuando se trabajó con personas o individuos y textos, a continuación se explica y justifica la particularidad de cada caso:

1. Todos los profesores del departamento de Biología de la Facultad de Ciencias Agropecuarias de la UCLV. Es importante aclarar que aunque el estudio se realiza con toda esta comunidad, solo 10 individuos de ella realizan los resúmenes para el proceso de evaluación del modelo.
2. Todos los Lingüistas con calificación de Catedrático o Profesor Titular con calificación de: Expertos en el desarrollo del idioma. Se selecciona esta población completa por su elevado nivel científico y metodológico.
3. Psicólogos y Pedagogos: Profesionales que se dedican a la actividad de la Psicología y la Pedagogía, intervienen como observadores de estrategias cognitivas.
4. Todos los bibliotecarios de la Universidad Central "Marta Abreu" de las Villas dedicados a la realización de resumen documental.
5. Usuarios Generales son todas aquellas personas que no son expertos en el tema, en esta comunidad se utilizaron 10 usuarios seleccionados al azar.
6. Todos los documentos que componen tanto el corpus inglés (4000 documentos) y el español 3250. Se utilizó como única muestra 50

documentos en español para el estudio del discurso del dominio Ingeniería de Puertos y Costas (Dominio Base), de estos 50 documentos se trabajó con 10 siempre que fue necesario desarrollar reglas de inferencia y análisis más precisos.

7. Todos los Especialistas en Ingeniería de Puertos y Buques de sobrada experiencia en el terreno de la investigación, una población seleccionada a partir de la tenencia de más de 25 años de experiencia en la labor (Especialistas).
8. Expertos en ORRIC (Organización Recuperación y Organización del Conocimiento). Diversos especialistas con determinados años de experiencia en la praxis del ORRIC y en la docencia.

Dentro de los sujetos que se vinculan directamente a las especialidades que se describen están los siguientes:

- Lic. Maricela Valdés Rodríguez
- Lic. María Nela Sibat Delgado
- Lic. José Monteagudo Fortín
- MSc. Manuel Osvaldo Machado Rivero
- MSc. Manuel Ernesto Ruíz Ramos
- Lic. Mayumis López Siverio
- Lic. Maydolis Agüero
- Lic. Dani Domínguez Pérez
- Dr. Luis Alfaro Echevarría
- Dra. Gema Mestre Varela
- MSc. Michel Faife
- Dr. Heriberto Expósito Santana
- Dr. César Chagoyén Méndez

- Dr. Armando Velásquez Rangel
- Dr. Angel Horta Amaro

A continuación reflejaremos en una tabla (tabla 1) la distribución de las comunidades de dominio para cada técnica de investigación.

| Técnica | Comunidad | Cantidad |
|---|---|---|
| Cuestionario | Ciencias Biológicas | 42 |
| | Especialistas | 5 |
| | Usuarios Generales | 10 |
| Observación | Ciencias de la Información | 12 |
| | Especialistas | 5 |
| | Lingüistas | 2 |
| Análisis de Contenido | Textos de Ingeniería de Puertos y Costas (Dominio Base) | 50 |
| Entrevista en Profundidad | Psicólogos | 6 |
| Cuestionarios | Expertos en ORRIC | 11 |

Tabla 1: Resumen de las comunidades de estudio

### 1.1.2.- Variables, operacionalización

El vocablo variable ha sido descrito en varios repertorios léxicos, una de las definiciones más sobresalientes y sencillas aparece en el diccionario Espasa (Calpe, 2008), donde se define a una variable como la capacidad para variar o para tomar diversos estados. Las variables salen del problema de investigación y, en este caso, las variables son: Modelo sociociocognitivo y Extracción y desambiguación de textos científicos. A continuación pasaremos a su definición conceptual y a su operacionalización.

### Definición Conceptual de las Variables

La definición conceptual de las variables radica en el posicionamiento conceptual de los valores de las variables según Kerlinger (Kerlinger, 2006). Las variables se han definido de la siguiente forma:

- Modelo sociociocognitivo: Conjunto ideal de operaciones que realiza un ser humano en un proceso dado.
- Extracción y desambiguación de textos científicos: Proceso que lleva implícito el acto de resumir y de dar sentido pragmático al texto mediante la cohesión, la coherencia y el balance textual.

**Definición Operacional**

La operacionalización de las variables en esta investigación consistió en la determinación de los valores y las dimensiones hacia las que se inclina la investigación. La tabla (tabla 2) refleja las variables, las dimensiones y los indicadores que miden las variables.

| Variable | Definición Operacional | Indicadores | Dimensiones |
|---|---|---|---|
| Modelo sociociocognitivo | Grado percibido en el modelo de las características del dominio Ingeniería de Puertos y Costas | Adecuación de las característica del dominio a las del modelo | Adecuación cognitiva |
| Extracción y desambiguación de textos científicos | Percepción de la calidad de la desambiguación y extracción de texto a partir de la adecuación de los textos extraídos a la realidad comunicativa que comunica (extracción | Percepción en la extracción y desambiguación de textos, de los elementos que distinguen al domino discursivo Ingeniería de Puertos y Costas | Calidad de la extracción y desambiguación |

Tabla 2. Comunidades de Práctica que participan en la Investigación

### 1.1.2.1.- Exposición de los principios de trabajo seguidos en la investigación

Esta investigación tiene como objeto la creación de un modelo para la extracción y desambiguación de textos científicos a partir de las perspectivas de la Ciencia de la Información. Las rutas de análisis que se defienden en este trabajo

asumen un estudio teórico de las formas de extraer textos, formulando análisis críticos sustentados en los retos que impone la Ciencia de la Información como disciplina instrumental. El modelo parte de la base del estudio del comportamiento de los seres humanos a la hora de resumir, siguiendo el proceso descrito por Endres-Niggemeyer (Endres-Niggemeyer, 2005) y mejorando las reglas de discurso desarrolladas por D 'Cunha (D'Cunha, 2006).

Este tipo de modelo posee también procesos de análisis lingüísticos amparados en pasos accesorios como: construcción de ontologías y construcción de base de datos. Para implementar el modelo se tomó como base el domino de la Ingeniería de Puertos y Costas, un espacio terminológico muy variado. Para ello se estudiaron los discursos inherentes a este dominio y se generaron reglas de discurso que fueron implementadas en el sistema, denominado PUERTOTEX. Finalmente se evaluó el sistema y el modelo propuesto aplicando diversos test y técnicas que se describen en detalle más adelante.

### *1.2.-Objeto de Investigación*

El objeto o escenario investigativo particular de esta investigación es el resumen automático, lo que representa el abordaje de una parcela de la Representación de la Información y el Conocimiento que viene sufriendo cambios en sus concepciones técnicas, metodológicas y procedimentales. En la medida en que se van desarrollando los procederes de tratamiento textual se distingue que los derroteros tradicionales de este campo se amplían hacia el uso de herramientas semánticas, análisis de dominios, tratamiento morfológico, redes de asociación y agentes, lo que supone el desarrollo multidisciplinar para este tipo de representación informativa que, según Hernández (Hernández, 2007), es una concurrencia de instrumentos particulares, en los espacios de producción, en el universo y complejidad de los objetos de análisis y en la gama de profesionales involucrados en este campo. La adecuación y la estructuración de los saberes que engloban hoy al resumen automático constituyen una suma de disciplinas sin las que sería imposible formular cualquier modelo de resumen automático.

### 1.2.2.1. – Campo de Investigación

En los estudios de Urías (Urías, 2009) se aclara que el campo de la investigación es el segmento particular más amplio donde cabe el objeto de la investigación, es el campo quien contiene al objeto y suele ser más amplio que este. El escenario investigativo global es la Representación de la información, un segmento de la Ciencia de la Información y la Bibliotecología que interactúa con diversos saberes como: la Inteligencia Artificial, La Ciencia de la Computación y la lingüística, disciplinas que conforman el escenario de análisis de este modelo.

Esta investigación se sustenta en una convergencia de conocimientos que conforman las herramientas de la Ciencia de la Información.

### 1.2.2.1.2.- Tipo de Investigación

Cuando nos referimos a tipo de investigación en la literatura de la Ciencias Sociales se declara la postura que se asume al investigar y qué es lo que se logrará con en el estudio en cuestión de acuerdo a la novedad del tema, los niveles de aplicabilidad, etc. (Urías, 2009). Estas cuestiones hacen que la investigación que se presenta a continuación se lleve a cabo con un corte descriptivo, teniendo en cuenta que en ella se describen los procesos que se utilizan para la confección de extractos de un único texto, así como la valoración del desempeño de las herramientas utilizadas en la confección de resúmenes, detallando finalmente un modelo para lograr dichos objetivos desde la Ciencia de la Información.

## *1.3.- Perspectiva de la Investigación*

Las perspectivas de las investigaciones toman varias direcciones acorde a las posiciones que va abordando el estudio en cuestión, según Baptista (Baptista et al., 2005) las perspectivas de las investigaciones pueden ser de corte cualitativo, orientada a las cualidades esenciales de los fenómenos que se investigan; cuantitativas, desarrolladas para estudios de naturaleza netamente estadística López Yepes (Lopez Yepes, 2005). En el caso de esta investigación se acogió

un basamento de corte mixto (cuantitativo y cualitativo) que se desempeñó fundamentalmente a través de:

- una mirada socio-constructivista para el descubrimiento, exploración y mejor comprensión de los saberes inherentes al Resumen Automático alejados de las perspectivas sustentadas en la Ciencia de la Computación y en la Lingüística clásica, posición que ha sido hasta el momento objeto de las mayoría de los estudios existentes en Ciencia de la Información sobre el tema, es decir el resumen automático es una entidad influida por la descripción de fenómenos netamente cibernéticos con una ligera parcela lingüística. Esto deviene una discusión entre la ortodoxia militante del procesamiento de la información textual en Ciencia de la Información y otras teorías emergentes, en las cuales el texto científico es el resultado de un fenómeno social e histórico que se reelabora constantemente.
- una valoración crítica sobre la generación de resúmenes automáticos centrada en sus basamentos teóricos, modelos, herramientas procedimentales y su formas de construcción dinámica, de forma que pudiera obtenerse un nuevo modelo de extracción sustentado en las deficiencias de modelos precedentes,
- un enfoque desde el pensamiento complejo declarado en los diversos análisis que enuncian la necesidad de considerar nuevas visiones en las modelizaciones para la construcción de resúmenes,
- una perspectiva sistémica viable, teniendo en cuenta las transformaciones modélicas a partir de la lógica sistémica, que redunda en la implementación de un sistema bajo la posición hegemónica de las comunidades de práctica,
- la defensa del análisis de dominio como exploración de regularidades para comprender los contextos en que se insertan los grupos sociales, que deben estar reflejadas en los modelos de Resumen Automático,

- la validación del análisis del discurso en el terreno de la Ciencia de la Información como herramienta del análisis de contenido capaz de reconocer y representar las características del dominio de conocimiento Ingeniería de Puertos y Costas.
- utilización del análisis documental clásico para desarrollar otras lecturas del fenómeno resumen automático, que intenten rebasar la dimensión hipercriticista y escéptica que ha reinado en los textos que desde la BCI (Biblioteconomía y Ciencia de la Información) han abordado este tema.
- uso de métricas de evaluación que cualifiquen la calidad de las ontologías y del sistema, teniendo en cuenta los principios heurísticos y las valoraciones realizadas para estos fines, dotando al modelo de una evaluación consistente y adaptada a las particularidades de la herramienta que se implementa,
- la cuantificación de los elementos retóricos, estilísticos y sintácticos para obtener una frecuencia que permita determinar reglas de carácter cognitivo que reflejen la diversas formas en que resume un ser humano,
- la obtención matemática de vocablos mediante herramientas lexicométricas y de evaluación de la calidad del léxico, elementos que prueban la necesidad de estudiar el dominio y su contrastación empírica y matemática,
- el uso de herramientas de cuantificación de n-gramas como una posición analítica que permita la comparación de los elementos topológicos de un texto, dando a la investigación una mirada clásica para la evaluación de los extractos resultantes de la aplicación del modelo,
- la valoración a la luz de la pragmática de los resultados del modelo por colectivo de jueces encargados de definir la calidad de los abstractos y del modelo en sí.

Estas dimensiones que justifican la investigación necesitan también de métodos de investigación lo suficientemente coherentes para desarrollar de forma eficaz

el proceso investigativo según apuntan Guadarrama y Urías (Guadarrama, 2009, Urías, 2009).

### *1.4.- Métodos de la investigación*

La definición conceptual de método viene dada como “modo de decir o hacer con orden una cosa” (Calpe, 2008). Otras conceptualizaciones sobre el concepto de método han sido descritas en Hernández (Hernández, 2007) , donde se establece que se trata de un conjunto de normas y procederes complejos para la realización de determinada actividad. En Baptista, Urías, Marimón Carrazana y Guadarrama (Baptista et al., 2005, Urías, 2009, Marimón Carrazana, 2005, Guadarrama, 2009) pueden verse algunas definiciones que rondan estas perspectivas o sentidos del término método:

- El sentido filosófico: "está constituido por el conjunto de operaciones intelectuales por las que una disciplina trata de alcanzar las verdades que persigue, las demuestra y las verifica" (Método en el sentido general de procedimiento lógico).
- Una actitud concreta en relación con el objeto: "dicta especialmente formas concretas de enfocar y organizar la investigación, pero esto de forma más o menos imperativa, más o menos precisa, completa y sistematizada" (ej. Método experimental, método clínico).
- Una tentativa de explicación, "se vincula más o menos a una posición filosófica /.../ y ante todo persigue un esquema explicativo que pueda ser más o menos amplio y situarse en un nivel de profundidad muy diferente" (ej.: el método dialéctico es empírico y supone observaciones concretas; el método estructural, etc.).
- Un dominio particular "se aplica a una esfera específica y supone una forma de proceder que le es propia (el método histórico, el método psicoanalítico y otros, también se relacionan igualmente con una concepción teórica de conjunto de la psicología o de la sociedad...).

El método de investigación es una forma regularizada, sistematizada y ordenada de realizar alguna acción para Guadarrama y Meneses Placeres (Guadarrama, 2009, Meneses Placeres, 2010). Estos autores refieren la existencia de balances teóricos entre las características esenciales de los métodos científicos, coincidiendo en que el método de investigación tiene las siguientes características:

- Trasciende los hechos.
- Es fáctico.
- Se atiene a reglas metodológicas.
- Se vale de la verificación empírica.
- Es autocorrectivo y progresivo.
- Sus formulaciones son de tipo general.
- Es objetivo.

Los métodos científicos que se utilicen en la investigación deben estar en concatenación con el objeto de estudio y el campo de investigación, por ello en esta tesis se apela a métodos de nivel empírico, matemático y nivel teórico en consonancia con las indicaciones que formulan tanto Urías como Rubio y Varas (Urías, 2009, Rubio and Varas, 2004).

### 1.4.1.- Teóricos

Los métodos de información de orden teórico son aquellos que se desarrollan a partir del conocimiento del la percepción del ser humano, y se basan en acciones lógicas del pensamiento Urías (Urías, 2009).

**Lógico-histórico**: Se emplea para estudiar el desarrollo lógico e histórico de los principales criterios sobre el tema. Se parte de la ubicación de los resúmenes científicos en la década del 60 del siglo XX. Es utilizado para secuenciar el desarrollo del tema en todos sus aspectos teóricos y metodológicos

**Analítico-sintético**: Posibilita analizar por partes los principales documentos y

consideraciones que describen el nacimiento y evolución del resumen científico hasta llegar a su actual proposición. Permite arribar a conclusiones sobre las metodologías estudiadas y el desarrollo del software.

**Inductivo-deductivo**: Los elementos particulares del resumen automático se toman como referencia, a través de la inducción analítica, para desarrollar la investigación. Se parte de las potencialidades de estos tipos de construcciones documentales para determinar los elementos que necesita para su uso en INTERNET.

### 1.4.2.- Empíricos

#### 1.4.2.1.- Técnica de Análisis documental clásico

A través de los métodos teóricos antes explicados se realizó un minucioso análisis documental, utilizando materiales escritos, sin perder de vista el análisis del contexto histórico, lógico y social. Se utilizó toda la información que brindó cada documento y se siguió la siguiente metodología para el análisis:

1. **Determinar los objetivos de estudio documental:** Dirigido a revelar los documentos que existen sobre el tema para realizar, mediante ellos, un trabajo de valoración crítica que refrende este tipo de análisis.
2. **Establecer la muestra de los documentos que serán estudiados:** Se establece una muestra de los documentos que serán analizados para determinar los elementos del resumen. Entre ellos están obras impresas, publicaciones seriadas impresas y digitales, así como diversos artículos de algunos buscadores especializados como: Scirus y Google scholar sobre el tema y publicaciones especializadas
3. **Determinar las unidades de análisis en las que se fracciona el contenido para estudiar el documento:** Teniendo en cuenta el contexto histórico se analizan paso a paso los períodos de producción editorial representados en los documentos. Se analiza el origen de las primeras aportaciones sobre la temática y el valor de las mismas para la investigación.

4. **Elaborar las categorías de análisis:** Se realizó a partir de los conceptos necesarios para comprender la importancia de los resúmenes y el desarrollo de sus arquetipos metodológicos. Unidades de Análisis: Resumen, Resumen Hipermedial, Hipertexto, Metodología, Sistemas Lógicos, Resumen Estructurado, Minería de Texto, Evaluación de Resumen.
5. **Realizar el estudio documental registrando la información:** Se realizó el registro de la información siguiendo la lógica de trabajo expresada en los pasos anteriores, sin perder el análisis contextual, tecnológico y filosófico de esta temática, logrando una integración coherente de todas las ideas para, de forma armónica, entrelazar y sistematizar todos los referentes teórico metodológicos que han rodeado al resumen documental desde 1960.
6. **Valoración de la información obtenida:** Se realizaron valoraciones a partir de los presupuestos teóricos y metodológicos precisados, realizándose inferencias, argumentaciones, críticas y, finalmente, arribando a conclusiones sobre los elementos que debe contener un resumen científico para ser expresado en una nueva metodología.

#### *1.4.2.1.- Modelación*

Es un método que se aplica a las investigaciones que llevan consigo el desarrollo de software y herramientas de análisis de contenido (Urías, 2009). Se distinguen específicamente por establecer conexiones lógicas sobre todo cuando se desarrollan sistema basados en software y en entidades. La modelación permite desarrollar operaciones lógicas de abstracción, por eso en esta tesis se utiliza en el desarrollo de la aplicación que sirve para el uso del modelo a través de PUERTOTEX de Domínguez (DOMÍGUEZ, 2011c).

##### 1.4.2.1.1.- El enfoque de sistema

Utilizado para la concepción sistémica de la herramienta y del modelo, donde las acciones van de las más simples a las más complejas y facilitan el desarrollo de

actividades de forma armónica. Según Urías es una técnica que facilita la concepción sistémica de una herramienta de información o un sistema de conocimientos (Urías, 2009). Sus ventajas esenciales son las siguientes:

- Permite desarrollar estrategias de análisis desde lo más simple a lo más complejo.
- Es un proceso inherente a la construcción de modelos y sistemas de enseñanza.
- Integra la concepción sistémica del pensamiento.
- Permite añadir conocimiento y aprendizaje de forma organizada.
- Es un enfoque eslabonado con teoría cognitivas.

### 1.4.3.- Técnica Observación Ajena

La observación es uno de los procesos de más alto nivel de empirismo que se ha utilizado en la investigación. Es un método cuya validez sustenta aquellas investigaciones donde se requiere obtener información sobre el comportamiento de algún elemento o cuestiones relativas a habilidades cognitivas (Baptista et al., 2005). Entre sus grandes deficiencias está la necesidad de que los observadores estén entrenados, ya que pueden solapar a veces la calidad de las observaciones en se han declarado estas técnicas. A continuación se muestran las particularidades de la observación en esta tesis:

- Guía de observación de las estrategias cognitivas para el acto de resumen. En esta investigación toman observación de los 12 especialistas de las 12 bibliotecas de la UCLV en el acto de resumen con el objetivo de determinar los pasos que hacen para el desarrollo de textos extraídos de diversas publicaciones relacionadas con la biblioteca de la Facultad en que trabajan (Anexo 1). Para realizar la observación se confeccionó una sesión de trabajo cuya guía se encuentra en el Anexo 1. La guías de observación en esta tesis poseen Categorías de Observación y Subcategorías que deben ser contabilizadas por los que dirigen la sesión de observación. Las variables que han de tenerse en cuenta son proceso

de lectura, análisis del documento y representación. Cada variable posee subcategorías que han de contabilizarse para luego hallar la media de las observaciones. Este test requiere de un proceso de entrenamiento para que los observadores (profesionales entrenados para el estudio) puedan adecuarse a los requerimientos de la investigación, para el cual se desarrolló una entrevista del tipo grupo focal.

- Análisis de los Indicadores de la Ontología. La guía de observación de la Ontología posee 3 variables y 13 indicadores (Anexo 2). Cada indicador debe ser marcado en la herramienta de análisis. Los especialistas observarán la ontología a partir de la herramienta Protex (Domínguez, 2010), desarrollada adhoc para este proyecto. Irán asentando los indicadores en cada uno de los elementos del contenido analizado en las casillas establecidas a tales efectos en Protex (Domínguez, 2010). Las variables a analizar son: Indicadores léxicos, Indicadores de evaluación de la estructura sintáctica, Precisión en la recuperación de documentos, Indicadores para la evaluación de la estructura sintáctica (Anexo 2). Esta técnica será realizada por 5 especialistas con elevado nivel de conocimiento del tema Ingeniería de Puertos y Costas.
- Guía de Observación para los rasgos y cualidades de los textos en inglés y español. 2 expertos en lingüística valorarán los textos en inglés y español a través de las siguientes variables: similaridad estructural, similaridad de imagen, similaridad léxica, Similaridad de Contenido (Anexo 3).

### 1.4.4.- **Técnica de Análisis de Contenido**

El Análisis de Contenido es el método de investigación más utilizado en el capítulo 4 de la investigación, es una forma de estudio de las particularidades de los mensajes, tanto escritos, como sonoros, audiovisuales, etc. Hernández (Hernández, 2006, 2007) Este método investigativo tiene su origen en las primeras descripciones empíricas de conocimiento referido en los trabajos de Urías, y los de Harter y Busha (Harter and Busha, 1990, Urías, 2009) en el

análisis de modelos de contenido esgrimidos para diversos fines. El fin que se le da en esta investigación al análisis de contenido es la caracterización del contenido de los textos de dominio Ingeniería de Puertos y Costas. Esta elección puede traer confusiones, pues hay autores como López-Huertas (López-Huertas, 2008) que prefieren que los estudios referidos a texto, sean denominados como análisis de domino. El autor no discrepa en la denominación que aporta al autora sobre el análisis de domino, solo se limita a enfatizar que la técnica que sustenta su investigación no es análisis de dominio, es análisis de contenido utilizado como herramienta procedimental en la que se inserta el análisis de discurso, el que constituye una forma específica del uso del lenguaje, como una forma específica de interacción social. Así, el discurso se interpreta como un evento comunicativo completo en una situación social a partir de las teorías de Van Dijk (Van Dijk, 2004), donde se enfatiza que no constituye una técnica con un cuerpo teórico propio y que se encarga de:

- El texto del discurso y, a la vez, sus condiciones de realización (contexto)
- Las dimensiones cognitivas, sociales y culturales de uso del lenguaje
- Los diferentes niveles que lo componen (fonológico, sintáctico, semántico y pragmático).
- Los interniveles (estilo, recursos retóricos, la argumentación)
- Los modos del discurso (oral, escrito, planeado, no planeado, formal, informal)

En Hernández (Hernández, 2007) se describen estos modos de actuación, donde se destacan las escuelas que a continuación se enuncian:

1. Formalistas rusos (análisis literario partiendo de Sauussure).
2. Círculo lingüístico de Praga (Jakobson y su modelo de comunicación).
3. Estilística (Su precursor es Bally y su objeto de estudio los enunciados asistemáticos).

4. Glosemática (A partir de los estudios de Hjemslev se integra con la Etnología, Sociología y Psicología).
5. Gramática estructural (su representante principal es Harris, EE.UU., que se ocupa de los aspectos formales de la lengua, suprimiendo la semántica).
6. Tagménica (Escuela de Pike, más allá de la oración, el acto comunicativo en su totalidad).
7. Lingüística funcional británica (cuyas figuras son Firth, Hallyday, y Leech, se aplica en la poesía y en los anuncios comerciales).
8. Estructuralismo checo (representado por Kristeva y su perspectiva funcional, coherencia y progresión).
9. Estructuralismo francés (Esta corriente la representa Greimas, y se centra en los esquemas narrativos culturales de los textos antropológicos por la influencia de Lèvy Strauss).
10. .Lingüística generativo-transformacional (análisis de la presuposición).

Teniendo en cuenta estas cuestiones los elementos que analiza el AD en esta investigación son los siguientes:

1. Cohesión (enlaces formales externos, estructuración sintáctica).
2. Coherencia (enlaces semánticos).
3. Intencionalidad (pragmática, actitud y propósitos del productor).
4. Situacionalidad (marco interaccionista, contexto sociocultural).
5. Intertextualidad (relaciones con otros textos).
6. Capacidad Informativa (transmisión de entidades y situaciones en un proceso cognitivo y emotivo).

Al igual que el análisis lexicométrico, el análisis de discurso ha aparecido como una herramienta para especificar el contenido de los mensajes dentro del análisis de contenido. Desde la óptica de esta tesis las aplicaciones del análisis

de contenido han servido para estudiar las características del corpus de estudio en su forma sintáctica, morfológica, estilística, lo que aporta reconocer los elementos siguientes en el corpus:

- Cantidad de Oraciones.
- Elementos Retóricos (Estructura Textual).
- Verbos.
- Ejes semánticos.
- Estructuras verbales.
- Elementos del Contenido que se suprimen.
- Elementos del texto que se usan para el resumen.
- Posición de los vocablos en el texto.
- Elementos imprescindibles en el texto.

Para estudiar el corpus de documentos en el proceso de análisis de Contenido, fueron construidas escalas de análisis a través de una guía de estudio de contenido en la cual se declaran los elementos esenciales para determinar las unidades léxicas del artículo científico en Ingeniería de Puertos y Costas. Esta guía se aplica a una muestra de 50 artículos en español de una población de 1435, seleccionados aleatoriamente por el doctorando.

A continuación se declaran específicamente cuales son los elementos del texto de la tesis que han sido objeto de análisis de contenido:

- Declaración de elementos que indican relevancia e irrelevancia en los textos de los textos del corpus: Este análisis particular se aplicó a 50 artículos en español del Corpus de Puertoterm, para ello se estableció una guía de análisis que reflejase los elementos a cuantificar. Esta técnica se realiza con el objetivo de seleccionar aquellos elementos textuales que pudieran tener peso o no en las oraciones del resumen de caras a su implementación como reglas de selección en la herramienta

computacional. El proceso consintió en leer el documento y seleccionar los elementos que indicaban relevancia oracional y los que indicaban irrelevancia oracional (Anexo 4).

- Análisis de la estructura de los corpus, esta guía tiene el objetivo de detectar las deficiencias en el corpus español. La guía cuenta de tres variables para el análisis: Presencia del Resumen del Autor, Estructura científica (IOMRC), Claridad en la redacción técnica. Para ejecutar esta verificación textual se realiza la lectura de 50 artículos de ambos corpus y se registran las veces que ocurren las variables declaradas con anterioridad, posteriormente se cuantifican y se calculan sus valores porcentuales. Se debe hacer para cada texto un análisis individual (Anexo 5).
- Construcción de los ejes semánticos de los textos con el objeto de obtener la estructura organizativa de los corpus como clave de organización de la ontología. Para ello la guía de análisis de contenido se centra en las variables que se declaran a continuación: Procesos, Procesos Naturales, Procesos Artificiales, Agentes Naturales, Artificiales, Medidas, Instrumentos, Efectos y Causas. El estudio se aplicará a 50 corpus, el proceso para la ejecución de esta técnica consiste en la lectura del documento, la selección de los términos clave y la graficación de los elementos mediante la construcción de un mapa conceptual mediante cmaptoolls (Anexo 6).
- También el análisis de contenido ha servido para la contabilización de las veces en que aparecen las reglas sintácticas comunicativas. El objeto de este tipo de estudio también se centra en obtener la frecuencia en que ocurren determinados fenómenos estilísticos, sintácticos y comunicativos en los textos. Para desarrollar este proceder se establecieron un total de 11 regularidades que se muestran a continuación: Regularidades en las que se elimina un satélite discursivo, Regularidades en las que se elimina un núcleo discursivo, Regularidades en las que no se separa el satélite de

su núcleo, Regularidades en las que no se separan dos núcleos, Regularidades en las que se elimina un satélite discursivo relacionado con un elemento sintáctico, Regularidades en las que se elimina un satélite discursivo relacionado con un elemento comunicativo, Regularidades en las que no se elimina un satélite discursivo relacionado con un elemento comunicativo, Reglas que eliminan elementos sintáctico-comunicativos, Regularidades donde no se suprimen aquellos satélites discursivos de elaboración, Criterios de desambiguación léxica. Todas estas variables se analizan en cada texto de forma individual para determinar su posición y frecuencia, calculando luego los valores porcentuales y registrando la información en una tabla dispuesta para ello (Anexo 6).

- Verificación de la presencia de los segmentos textuales del texto fuente en el resumen. Para ello se analizan 50 textos de los que se estudian aspectos como: cantidad de palabras en los segmentos del texto, inclusión de segmentos del texto fuente en el resumen y oraciones que se utilizan tanto en el resumen como en el texto fuente (Véase 5.5.2.3), también se someten a verificación los verbos utilizados en cada uno de los textos con el objeto de determinar su frecuencia en los textos estudiados, para este último análisis se utilizaron 10 textos solamente.
- Verificación de la estructura de los corpus, una técnica implementada para seleccionar la calidad de los corpus y determinar cuál será el corpus base o patrón y el corpus que servirá de tratamiento. Para llevar a cabo este estudio se utilizan las ontologías wordned y eurowordned, las mismas se introducen en la herramienta de visualización de la herramienta Protex que además tiene dos plugin de carga de documentos, con las ontologías cargadas se someten los corpus a verificación y contrastación (Anexo 21).
- Estudio de las dimensiones estilísticas de los artículos. Técnica que consiste en el estudio de cada artículo científico de ambos corpus para

determinar si poseen las siguientes variables: artículos léxicos, artículos divulgativos y artículos científicos. Se debe registrar toda la información de cada documento y calcular el porciento de acuerdo a la cantidad de textos que contiene el corpus.

Estudio de los elementos cohesivos del texto, es decir determinación de la posición que en el texto tienen las estructuras sugeridas por Mann y Thompson (Mann and Thompson, 1990). Las variables que se analizan en los 50 Corpus son: Elaboración, Evidencia, Fondo, Justificación, Contraste, Unión, Secuencia, Lista, Concesión, Condición, Propósito, Reformulación, Resultado, Resumen e Interpretación. Para ello se parte de un análisis de los textos y la declaración en cada una de estas estructuras de sus núcleos y satélites. Estas distribuciones textuales son las formas que declaran la estructura interna del texto y sirven para determinar las reglas del discurso que se automatizan en el Capítulo 6 de la tesis (Anexo 20).

- Identificación de unidades léxicas correspondientes a las siguientes variables: Unidades léxicas nominales, Unidades léxicas verbales, Unidades léxicas del título principal. Para determinar estas unidades se leen los 50 textos y se declara la presencia de las variables en cada uno de ellos, luego se contabilizan los resultados en una tabla creada para recoger la información (Anexo 19).
- Con el objeto de reforzar los resultados obtenidos en la evaluación de los resúmenes mediante la aplicación de Rouge, se decide analizar el contenido de los textos mediante la valoración de los 2 expertos en lingüística. Las variables que analiza esta técnica de investigación son las siguientes: cohesión, coherencia y balance textual (Anexo 15).

Estas técnicas no fueron sometidas a prueba de pilotaje debido a que ya están validadas en los estudios de D'Cunha, Zaldua y Satriano y Moscoloni (D'Cunha, 2006, Zaldua, 2006, Satriano and Moscoloni, 2000).

### 1.4.4.- Técnica de Cuestionarios

El cuestionario es un instrumento consistente en un conjunto de preguntas en torno a una o más variables para que se desean medir según afirma Baptista (Baptista et al., 2005). Su capacidad para aplicación es efectiva para los estudios de grandes grupos de personas. El cuestionario está compuesto de preguntas abiertas y cerradas, que sirven para establecer opiniones y para obtener valores específicos de cada variable. En esta investigación se ha seleccionado el cuestionario específicamente por su capacidad instrumental en la obtención de datos para poblaciones con un tamaño considerable según Harter y Busha (Harter and Busha, 1990). El cuestionario como método de investigación puede ser aplicado a diferentes estudios como:

- Estudios de raiting.
- Estudios de mercado.
- Análisis de opiniones de Servicios de Información.
- Estudio de comunidades.

Una de las características esenciales del cuestionario reside en su capacidad para articular determinadas preguntas con nivel de codificación específico. En Harter y Busha (Harter and Busha, 1990) se evidencia que los cuestionarios deben estar confeccionados con escalas de valoración, indicadores de análisis y variables, además de la demanda de cooperación para los usuarios. Esta características se han visto refrendadas por los trabajos de Baptista quiénes comparten el mismo criterio que los autores que se han mencionado en este acápite (Baptista et al., 2005). Los cuestionarios de esta investigación se aplican en diferentes momentos. A continuación se explicará en qué momentos han sido utilizados para aclarar las poblaciones con las que se aplican.

#### 1.4.4.1.- Cuestionario para el diseño del Servicio de Información

Cuestionario para estudio de necesidades. Este instrumento se desarrolló analizando 5 variables y 11 preguntas abiertas. Se aplica 42 usuarios, toda la población del Departamento de Biología de la Universidad Central" Marta Abreu"

de las Villas. Las preguntas son todas cerradas y evidencian cualidades del servicio de resúmenes que se deben construir en su diseño. Incluyen elementos de construcción, frecuencia de servicio, formato de salida, así como variables de calidad de los textos para su evaluación. El cuestionario posee una demanda de cooperación, especificación de variables que se utilizan, las subcategorías de las variables y las formas en que debe contabilizarse el contenido de los documentos (Anexo 7). Las variables utilizadas están elaboradas a partir de los criterios de Núñez (Núñez, 2005).

#### 1.4.4.2.- Cuestionario de Evaluación de Ontología

Para evaluar la calidad de la Ontología antes de desarrollar el sistema se han tenido en cuenta el análisis de determinados parámetros que sirven para constatar los requerimientos del sistema (Anexo 8). La técnica de investigación utilizada para la validación de la evaluación de la ontología ha sido el cuestionario, utilizado para evaluar de forma heurística su capacidad operativa. Este tipo de test se aplica después de evaluar la ontología mediante la observación de 5 expertos en el tema de la Ingeniería de Puertos y Costas, en los estudios de Urías (Urías, 2009) se le llama cuestionario de definición o instrumento de estudio secundario. Este cuestionario se aplica a los 5 especialistas en el tema, entrenados en el uso de ontologías. Este instrumento posee solo una variable denominada: recuperación de información. Para desarrollar el cuestionario solo hay que buscar en las diversas clases de la ontología y localizar lo que se indica en las preguntas, declarando el resultado en la planilla la búsqueda de cada elemento debe hacerse 3 veces y reflejar en el modelo en qué medida se logra hacer lo que pide cada pregunta, para ello se deben llenar las casillas del documento del Anexo (Anexo 8) y establecer los valores en las mismas. Los referidos valores son: “si”, “a veces”,”nunca”.

#### 1.4.4.3.- Test de Usabilidad de Sistema

El test de usabilidad es un instrumento de evaluación heurística desarrollado para evaluar el sistema de resúmenes. En la evaluación de la usabilidad del sistema se han tenido en cuenta algunos parámetros entre los que se

encuentran: navegación, funcionalidad, control del usuario, lenguaje y contenido, ayuda en línea, información del sistema, accesibilidad, coherencia, prevención errores y calidad arquitectónica (Anexo 9). Los participantes en el test de usabilidad son los 42 profesores pertenecientes a la Comunidad Ciencias Biológicas y los 5 especialistas en el tema. En una segunda fase de esta técnica se le pide al usuario (Comunidad Usuario General) que realice 8 tareas de trabajo con el sistema. El espacio entre cada medición será de 20 minutos. Debe marcarse con 0 las tareas no cumplidas, con un 1 aquellas conseguidas con dificultades y con un 2 las que ha realizaron fácilmente. En este cuestionario no se declararon variables ya que la variable de uso es una sola, Recuperación de la Información. Las preguntas que deben contestarse son 8 y los resultados deben asentarse en la tabla diseñada para ello.

#### 1.4.4.4.- Evaluación de la Herramienta

Cuestionario de evaluación de sistema. Un cuestionario que mide: Navegación, Funcionalidad, Control del usuario, Lenguaje y contenido, Ayuda en línea, Información del sistema, Accesibilidad, Coherencia, Prevención de errores y Claridad arquitectónica. El test está compuesto por 10 preguntas cerradas cada una acorde a cada variable (Anexo 10) y pueden ser contestadas por el usuario interactuando con el sistema. Los que harán este tipo de evaluación son los 5 expertos y los 42 profesores del departamento de Biología de la Universidad Central" Marta Abreu" de las Villas. El cuestionario se responde realizando las tareas que se indican en la documentación, el valor de los resultados se coloca en la tabla desarrollada para esto, donde aparecen los valores siguientes:"si","no", "a veces".

Las preguntas que se han elaborado para todos los instrumentos son cerradas para lograr con ellas una misma respuesta y facilitar las valoraciones al final de cada instrumento, aspecto que aconseja Urías (Urías, 2009). A continuación se explica cada cuestionario utilizado en la tesis y su forma de aplicación:

### 1.4.4.5.- Criterio de Expertos

La acepción terminológica experto, se identifica con grupo de personas de elevado nivel del conocimiento en un tema determinado, capaces de ofrecer juicios sobre algún fenómeno o cosa en forma conclusiva, o hacer aportaciones y valoraciones de alto nivel. Como indica Meneses Placeres (Meneses Placeres, 2010) *es el investigador quien debe determinar las cualidades que han de poseer los expertos a los cuales va a someter su trabajo* (Anexo 11).

En esta investigación se aplicó el criterio de expertos a través de tres fases específicas:

- Identificación del coeficiente de experticidad de cada especialista escogido.
- Evaluación de la calidad del modelo.
- Emisión de Juicios de valor de la calidad del modelo propuesto.

Ninguno de estos cuestionarios es sometido a prueba piloto debido a que estas mismas variables de estudio han sido utilizadas en estudios debidamente citados en la investigación como los de Nielsen y Floria (Nielsen, 1994, Nielsen, 2002a, Nielsen, 2002b, Floria, 2000).

### 1.4.4.6.- Entrevistas de Grupos Focales

En este estudio se empleó la entrevista por grupos focales (Anexo 13), definido por Rubio y Varas (Rubio and Varas, 2004) como:

*una técnica que trata de captar la realidad social a partir del debate o la discusión en pequeños grupos (…) Se trata de reproducir aquello que sucede en la sociedad (macro situación), a través de un grupo de personas (micro situación) reunidas a propósito por el investigador para hablar sobre un tema su flexibilidad de aplicación,*

*y que* según Meneses Placeres (Meneses Placeres, 2010) se *basa en una serie de aspectos que el entrevistador tendrá en cuenta a la hora de entrevistar; pero que no está en la obligación de seguirlos en el orden en que fueron planteados,*

*teniendo, además, la posibilidad de readecuarlos en su formulación en correspondencia con las particularidades que adopte la situación comunicativa, pudiéndose aplicar a grupos de personas o colectivos.*

Para realizar la guía de observación de estrategias cognitivas fue necesario realizar entrevistas en profundidad con el objeto de lograr la discusión y análisis de los parámetros que debían incluirse o eliminarse en este instrumento, ya que el estudio que antecede esta investigación fue desarrollado por Endres-Niggeyer (Endres-Niggemeyer, 2005) en otro contexto. Cuando se aplicó la entrevista los psicólogos expusieron sus criterios sobre el procedimiento a aplicar y las formas en que debía construirse el instrumento. Según Rubio y Varas (Rubio and Varas, 2004), las características de esta técnica se resumen en:

- Su objetivo esencial se centra en la recolección de vivencias y experiencias de los entrevistados.
- Demandan un clima de confianza en su aplicación, cuestión esencial para que se active el sistema de comunicación y el acto investigativo fluya de acuerdo a lo que sus objetivos.
- La concesión de estructura de diálogo es vital en la entrevista, para activar con esta forma elocutiva los recuerdos, vivencias, emociones y también los errores cometidos.
- Es un juego de artes donde la posición del entrevistador siempre tiene que ser estratégica, logrando manipular al entrevistado hacia las áreas o aspectos a los que se centra la investigación.
- Ayuda al entrevistado a construir un discurso coherente y centrado en las demandas del objeto y la praxis de la investigación.
- Está estructurada de forma que el discurso que se emplea, se fácil de decodificar.
- Constituye un proceso de entendimiento tamizado, donde los interlocutores interactúan para lograr una comunicación con consenso.

Las entrevistas, pueden ser de diversos tipos para Horacio Saldaño (Horacio Saldaño, n.d.):

- *Estructuradas: son aquellas en que el entrevistador ha preparado un cuestionario muy detallado para obtener la información, normalmente se da cuando el investigador conoce mucho del tema y se encuentra en una fase de la investigación donde necesita detalles puntuales.*
- *No estructuradas: son aquellas utilizadas para obtener información acerca de una historia de vida, de hechos o sucesos donde el protagonista es el entrevistado (…) si bien la entrevista no estructurada no lleva preguntas definidas es importante tener guías de preguntas o ítems previamente establecidos, para que durante la conversación, tratar de tocarlos para obtener información.*

La planificación de la entrevista (Tabla 3) se realizó de la siguiente forma:

| Entrevistado | Cargo | Lugar de Entrevista | Fecha de Realización | Duración |
|---|---|---|---|---|
| Lic. Alibet Viera | Psicólogo | Departamento de Computación J.C. Remedios | 29/1/09 | 1-5 pm. |
| Lic. Ariel González | Psicólogo | Departamento de Computación J.C. Remedios | 29/1/09 | 1-5 pm. |
| MSc. Bárbara Fuentes Cabrera | Pedagoga | Departamento de Computación J.C. Remedios | 29/1/09 | 1-5 pm. |
| MSc. Aymé Duquesne Morell | Pedagoga | Departamento de Computación J.C. Remedios | 29/1/09 | 1-5 pm. |
| Dra. Inés María Mederos Morell | Pedagoga | Departamento de Computación J.C. Remedios | 29/1/09 | 1-5 pm. |
| MSc. María Aleida Hernández | Pedagoga | Departamento de Computación J.C. Remedios | 29/1/09 | 1-5 pm. |

Tabla 3: Conjunto de Observadores (Psicólogos y Pedagogos)

Las entrevistas se desarrollaron en una sola sesión y tuvieron como objetivo recolectar información sobre los indicadores siguientes:

- Estructuración de los Procesos Cognitivos.
- Calidad de los Procesos Cognitivos.
- Importancia de los Procesos Cognitivos para el Resumen automático.

No se desarrolló una prueba piloto ya que estos procesos fueron validados en el estudio de Endrez-Niggemeyer (Endres-Niggemeyer, 1995), por lo que esta entrevista grupal consistió en debatir la calidad del método desarrollado por la autora en la automatización de los procesos cognitivos relativos al resumen (Anexo 13). Los resultados de la entrevista se exponen en el (Anexo 14).

Este proceder se aplica a los psicólogos y se caracterizó por lo siguiente en esta tesis:

Objetivos:

- Determinar las posibles habilidades cognitivas que debe desarrollar en un resumidor.
- Establecer las variables necesarias para desarrollar la guía de observación a los resumidores.
- Indicar los puntos de evaluación y la forma de registro de la información para la guía de observación.
- Indagar sobre la importancia de las estrategias cognitivas en la automatización y la forma en que estas pueden desarrollarse.

Aplicación:

La aplicación de esta técnica fue en una sola sesión dirigida por un facilitador. Los temas debatidos fueron los siguientes:

- Cuáles son las habilidades cognitivas necesarias para el acto del resumen en un ser humano.

- Las estrategias cognitivas en la escritura son lineales o iguales en todas las personas de un mismo dominio.
- Qué puntos o elementos serían los más efectivos para la evaluación y descripción de estrategias cognitivas en el acto del resumen.
- Qué importancia le atribuyes al aprendizaje de las reglas o estrategias cognitivas en las comunidades de práctica.

### *1.4.7.- Métodos de Nivel Matemático*

Los métodos matemáticos utilizados en esta tesis van desde la estadística descriptiva más simple, hasta complejos análisis estadísticos y matemáticos sustentados en algoritmos de elevada complejidad computacional. Explicar en detalle en este capítulo inicial todas estas cuestiones matemáticas haría más engorrosa la exposición del informe, por ello el doctorando solo se limitará a justificar de forma general los métodos matemáticos utilizados, pues, dentro de cada capítulo se explican en detalle. Las técnicas utilizadas en el desarrollo de la tesis fueron las siguientes:

#### 1.4.7.1.- Análisis Porcentual

Este es uno de los procederes matemáticos más utilizados en la estadística, el mismo referencia en que media se cumple determinada cuestión según se describe en Urías y Cué (Urías, 2009, Cué, 1988). Mediante el análisis porcentual se han cuantificado y valorado determinados test en la investigación, entre los que se encuentran el análisis de contenido, el estudio de usabilidad, la evaluación de la Ontología y el Estudio de Necesidades, acciones que se declaran en el capítulo 5 y en el 7. En el estudio de la ontología se calculan diversos indicadores utilizando porciento mediante el uso de la herramienta Protex de Domínguez, (Domínguez, 2010), los indicadores que se analizan son los siguientes: indicadores léxicos, indicadores para la evaluación de la estructura sintáctica e indicadores de recuperación de la información. El cálculo se realiza declarando en la interfaz de la herramienta Protex el indicador y la forma en que será evaluado.

### 1.4.7.2.- La Media

Se usa en la tesis para obtener la media en el caso que las categorías de análisis en la observación tengan sub-elementos de estudio. También en nuestra investigación se ha declarado la media para registrar el promedio de veces que una persona hace una acción cognitiva. La media es un método muy eficaz para el estudio de repeticiones de diversos valores Cué (Cué, 1988) y es susceptible de usarse cuando la población a observar es muy pequeña.

### 1.4.7.3.- Rouge

Para valorar la eficacia del modelo en comparación con otros modelos se utilizó Rouge, una herramienta de valoración matemática de resúmenes desarrollado por Lin y Hovy (Lin and Hovy, 1998). El cálculo de Rouge consiste en la valoración de la similitud de los n-gramas de un texto base contra un texto construido con métodos automáticos. El cálculo de n-gramas tiene variaciones que determinan la profundidad de la coincidencia de los textos comparados, las variantes que se han utilizado en la investigación son las siguientes: Rouge-N, Rouge2, Rouge-SU4, RougeL y RougeW. En esta tesis el análisis con Rouge se realiza contra resúmenes realizados por modelos de extracción de microsoftword, una colección baseline, resúmenes hechos por los expertos en el tema y las reglas puertoterm (puertot-). El uso de estas variantes de cálculo facilita el análisis de estos textos desde el punto de vista geométrico y la similitud topológica, no así en el orden la de la coherencia y el balance textual Senso y Leiva (Senso and Leiva, 2008). Muchos textos valorados por Rouge son similares desde el punto de vista topológico, pero muy bajos en calidad pragmática, esto quiere decir que tener en el texto niveles elevados de coincidencia en Rouge, no implica que éste pueda leerse, es por ello que hay que hacer siempre una valoración por parte de los jueces de la coherencia, del balance textual y de la cohesión alcanzada en la extracción, solo así será evaluado un texto en su sentido utilitario. Rouge se aplica mediante Análisis de 10 resúmenes realizados utilizando las técnicas de abstracción y extracción. En este segmento de la investigación se le pide a 10 especialistas del grupo

Ciencias Biológicas que realicen resúmenes utilizando las técnicas referidas, las que servirán para evaluar los modelos de resumen en el capítulo referente a evaluación(Anexo 25).

### 1.4.7.4.- Técnicas para determinar la Calidad del Léxico

Múltiples son los métodos de agrupamiento que se utilizan en esta tesis, es por ello que se han seleccionado muchas variantes para calcular la calidad de los términos en los grupos textuales teniendo en cuenta las cualidades de la investigación. Las variantes de cálculo utilizadas para evaluar el léxico base en esta tesis están descritas en el capítulo 7, donde se explica su función y el criterio de selección de cada una, a continuación se listan las métricas que se utilizan en la tesis para evaluar el texto: Palabras temáticas importantes, Umbral de frecuencia de términos y Ley de ZIF presentada por Bolelli (Bolelli et al., 2007), Umbral de frecuencia de documentos, Razón de señal a ruido (Signal-to-Noise Ratio), Calidad de términos, Skewness y Kurtosis, Overall Simillarity y F-Measure que aparecen en los trabajos de Arco (Arco, 2008). Para construir los grupos se ha aplicado el algoritmo VSM (Vector Space Model) desarrollado por Salton y su colectivo (Salton et al., 1975) dentro de una herramienta denominada FOXCORP elaborada por Domínguez (DOMÍGUEZ, 2011b), otro de los algoritmos de agrupamiento que se utiliza en el agrupamiento del corpus es el utilizado en corpus Miner de Arco (Arco, 2005) el cual permite mostrar la información de aquellos resúmenes que carezcan de superestructura.

### 1.4.7.5.- Chi Cuadrado

**Chi cuadrado ($X^2$)** Es un estadígrafo establecido para estudiar hipótesis y relaciones entre variables u observaciones (Bécue, 1997, Cué, 1988). En este estudio las variables son siempre cualitativas, ya que declaran las cualidades de un objeto. Cada variable calculada con este estadígrafo tiene dimensiones específicas, por ende, de acuerdo a sus características y los resultados Chi cuadrado es una distribución de probabilidad continua con un parámetro k que representa los grados de libertad de una variable aleatoria. Las aplicaciones esenciales de este estadígrafo están en independencia de los estudios que se

realicen, además permite calcular la bondad de ajuste y la inferencia estadística. En esta investigación se ha aplicado esta medida estadística para estudiar la varianza de las observaciones de 2 lingüistas en el corpus inglés y el español a un 0.05 rango de error. La técnica consiste en registrar por pares las observaciones en una tabla elaborada para la prueba de observación descrita en el acápite Observación.

### *1.4.5.- Etapas de la Investigación*

La investigación que se propone tiene diversas etapas que se irán declarando a continuación. Al decir de Urías y Meneses Placeres (Urías, 2009, Meneses Placeres, 2010) las etapas de investigación son pautas que van desarrollándose en el proceso investigativo, teniendo en cuenta el tiempo y las condiciones de los investigadores. La Investigación se realizó en tres fases: fase 1 estudio de las técnicas y procedimientos asociados al resumen documental, fase 2 estudio empírico de las comunidades de práctica y fase 3 evaluación e implementación del sistema. Cada fase posee subfases que facilitan la estructuración de la actividad investigativa.

La **fase 1** consistió en la identificación de los referentes teóricos metodológicos de la investigación, haciendo énfasis en el campo del resumen documental, la misma se estructura en determinadas subfases:

- **Subfase 1:** Determinación de los documentos del estudio documental y la supresión de documentos de bajo nivel de aportaciones teóricas.
- **Subfase 2:** Identificación de los elementos metodológicos que sustentan el resumen documental en su forma automática.
- **Subfase 3:** Selección de la teoría y los sustentos metodológicos para construir el modelo de resumen.
- **Subfase 4:** Aplicación de la guía de observación para determinar las estrategias cognitivas de los resumidores.
- **Subfase 5:** Elaboración del modelo de resumen.

La otra parte de la investigación se desarrolló a partir de la **fase 2**, que no es más que el estudio empírico de las comunidades de práctica para la construcción del modelo. Las subfases que se inscriben en esta fase son las siguientes:

- **Subfase 1:** Aplicación de guías de observación con el fin de determinar las estrategias cognitivas de los resumidores de las Bibliotecas de la Universidad Central" Marta Abreu" de las Villas.
- **Subfase 2**: Estudio de las características del discurso Ingeniería de Puertos y Costas, mediante análisis de contenido para la construcción de reglas discursivas basadas en la cohesión del texto.

Para el final de la investigación se realizó la **fase 3,** implementación y evaluación del modelo. Los períodos o fases que corresponden a la fase tres se exponen a continuación:

- **Subfase 1:** Construcción de la documentación del sistema, diagramas de UML para el desarrollo planificado del software.
- **Subfase 2:** Implementación de las reglas textuales en Python utilizando la estrategia establecida en la fase 2 de la investigación.
- **Subfase 3:** Estudio de necesidades de información para desarrollar el servicio de resúmenes a través de cuestionarios.
- **Subfase 4:** Construcción y Evaluación de la Ontología para dar soporte al sistema. En esta fase de desarrolló el acopio de los términos y la estructuración del conocimiento tomando como base el patrón predefinido por el proyecto Puertoterm.
- **Subfase 5**: Aplicación de técnicas matemáticas de evaluación de la calidad del modelo. Aplicación de Rouge y Técnicas de Selección de Términos y $X^2$ (Chi cuadrado).

- **Subfase 6**: Desarrollo de técnicas empíricas para la evaluación del sistema de resúmenes (aplicación de test de usabilidad de sistema a los usuarios y a los expertos en bibliotecología).
- **Subfase 7:** Evaluación mediante criterio de expertos de la calidad del modelo propuesto (Cuestionario de Evaluación Final).

### *1.4.6.- Herramientas de Programación y Modelación*

Las complejidades de la investigación obligaron a que el doctorando tuviera que utilizar determinadas herramientas de programación y de modelación para asumir la etapa de implementación del modelo, estas herramientas sirvieron para estructurar textos, confeccionar las bases de cocimientos y hacer la ontología. Entre las herramientas que se utilizaron están las siguientes:

**Protégé:** Sistema para edición de ontologías, facilitó la construcción de Ontosatcol (ontología del sistema).

**FOXCORP:** Herramienta para la evaluación del corpus del proyecto Puertoterm desarrollada a la medida por Domínguez (DOMÍGUEZ, 2011b) para evaluar las medidas de calidad de los corpus, estudio que permite reconocer cuál es el corpus más fuerte para que sirva de patrón para la construcción del sistema léxico y cuál el de menor capacidad semántica para ser tratado en la implementación del sistema.

**Python:** Lenguaje de programación orientado a objetos que posee una gran capacidad para trabajar con ontologías.

**Word Smith Toolls**, herramienta de análisis lexicométrico, capaz de facilitar la contabilización de los términos y sus frecuencias de aparición en el corpus.

**Satcol:** herramienta con una capa de agentes capaz de reproducir las estrategias cognitivas de los resumidores, desarrollada por el Ing. Sandor Domínguez Velasco (DOMÍGUEZ, 2009).

**TEXMIX:** De (Domínguez, 2011a) herramienta desarrollada dentro de FOXCORP (Domínguez, 2011 a) para formular correspondencias entre términos.

Funciona como un diccionario de la lengua con relaciones de meronimia, hiponimia, hiperonimia, meronimia, etc.

**Coculuscorpora:** Permite conocer la calidad de un corpus a partir de técnicas estadísticas, que valoran la calidad de los términos. También fue desarrollada por Domínguez Velazco (DOMÍGUEZ, 2011a).

### *1.4.7.- Análisis de las Fuentes de Información*

Todas las fuentes informacionales utilizadas en la tesis tienen un origen polisémico y sustentan cada una de las aseveraciones declaradas en esta tesis. La complejidad de las fuentes y del tema hizo que constantemente el autor tuviese que recolectar información de diversos recursos y fuentes. El procedimiento que se desarrolló para el análisis de fuentes consistió en lo siguiente:

1. Localización de Información en diversas bases de datos y recursos de información de diversa naturaleza.
2. Selección y análisis crítico de los temas y los documentos.
3. Confección de una biblioteca digital con la herramienta Endnote versión XIII.

#### 1.4.7.1.- Recursos de Información utilizados en la tesis

Diversos y abundantes han sido los recursos de información utilizados en esta tesis, debido a lo polisémico del tema, a los diversos dominios que se insertan en él y a la preponderancia de los estudios que sobre Procesamiento del Lenguaje natural ha existido en estos últimos años.

Los recursos de información utilizados en la investigación fueron los siguientes:

- ***Obras de Referencia*:** Básicamente se han utilizado la Enciclopedia of Library of Information Science, donde aparece un trabajo de Hjorland. El diccionario Alonso (Alonso, 1958) donde se describen algunas cuestiones relativas al uso de algunas palabras en la lengua española. También se utilizaron muchos tutoriales sobre Python, RDF, SPARQL, MYSQL y XML.

- ***Monografías y Revistas Científicas Impresas:*** Todas las Revistas que se utilizaron se encuentran en la biblioteca de la Facultad de Comunicación y Documentación de la Universidad de Granada, sin embargo, su uso fue bastante bajo ya que la mayoría de la información que se utilizó en la tesis fue netamente on-line o en formato pdf.
- ***Bases de Datos Científicas en Línea:*** La mayoría de los textos provienen de bases de datos que permiten el uso de de los textos completamente (a texto completo) ellas son:
- ***Web of knowledge*****:** Base de datos de elevado prestigio cuya cobertura temática abarca más de 250 ramas de las Ciencias aplicadas y de las Ciencias Puras, las Artes y las Humanidades. En esta base de datos existen alrededor de 8.5000 revistas, lo que representa alrededor de 1,4 millones de artículos, además incluye cerca de 2000 monografías y más de 4000 recursos a texto completo.
- ***Science Direct*****:** Base de datos contentiva de un 27 % de la literatura científica, tecnológica y de ciencias de la salud a nivel internacional. Oferta, libros, monografías, manuales, obras de consulta y más de 82 millones de artículos científicos.
- ***ACM Portal*****:** Un sitio de bases de datos, cuya especialización es la Ciencia de la Computación, posee innumerables recursos como: Revistas, Boletines de Noticias y Proceeding de Eventos.
- *Scopus*: una base de datos que permite el acceso a más de 3,4 millones de resúmenes de artículos, facilitando el contacto con los proveedores de aquellos recursos que aparezcan en la base de datos y sean a texto completo.
- ***E-LIS (Open Archive fron Library and Information Science:*** Repositorio que facilita el acceso a más de 7000 recursos especializados en materia de Bibliotecología y Ciencia de la Información cuyos derechos de edición están cedidos a este recurso por sus usuarios.

- ***Recursos Web*:** Las páginas del World Wide Web Consortium (W3W) han servido para el estudio de las recomendaciones para el desarrollo de la aplicación.
- ***Tesis y Documentos No Publicados:*** 15 Tesis sobre el tema desarrolladas en el terreno de la Ciencia de la Computación aportaron nuevas actualizaciones a los paradigmas de este tema, al igual que 3 Manuales y 7 informes de investigación terminada.
- ***Libros:*** Facilitan el análisis de los referentes de muchas de las teorías que se declaran en esta investigación.

El estudio de fuentes de información

### 1.4.7.2.- Resultados del Análisis de las fuentes de Información

La búsqueda en base de datos de nivel internacional se basó en el período 1982-2010. El primer criterio de búsqueda fue el término Tex Mining, obteniendo como resultado un recobrado 269 registros, este resultado es pobre si se tiene en cuenta la cantidad de trabajos que existe a nivel mundial sobre este tema. Este resultado hizo que hubiese que ajustar la búsqueda a las condiciones del fenómeno que se estudia, por lo que fue necesaria la búsqueda por el término *Summarization,* recuperándose 2269 registros. Concluido este paso se seleccionaron los indicadores de productividad que reflejaran las características de la producción en el tema [autores, instituciones, fuentes, años de publicación, relación disciplinar y colaboración entre autores] con el objeto de determinar dónde se centra la producción el tema, las fuentes más importantes y los autores más citados para identificar los autores más productivos en esta materia.

#### 1.4.7.2.1.- Autores más citados y Revistas

Los 20 autores más citados en el tema de Summarization se muestran a continuación. Inderjet Mani, autor con notables trabajos en el terreno de la Minería Textual, la Summarización de documentos, la Recuperación de la Información, la Minería de Datos y la Construcción de métodos para extracción de texto. Las aportaciones de Mani, rebasan los terrenos meramente

metodológicos, pues son varias sus contribuciones donde se tratan temas de implementación práctica. A Mani le sigue Kupiec, autor de importantes contribuciones en el terreno de los algoritmos para la sumarización y extracción automática de textos, también con varios trabajos prácticos de elevada utilización. Se demuestra en este análisis que no existe ningún autor del terreno de la Ciencia de la Información o de la Documentación entre los más productivos, sin embargo sí aparecen citados en el lugar 80 trabajos de María Pinto Molina y Wilfried Lancaster sobre la interdisciplinariedad de la extracción de texto, siendo los autores más representativos en nuestro dominio (Ver Tabla 4). La revista más citada es Automatic Summarization, especializada en el tema del resumen automático y ANN INT ACN SIG, debido a que los autores más citados publican en ellas.

| No. | Citas | Autores |
|---|---|---|
| 1 | 28 | MANI I, 2001, AUTOMATIC SUMMARIZAT |
| 2 | 26 | KUPIEC J, 1995, P68, P 18 ANN INT ACM SIG |
| 3 | 23 | EDMUNDSON HP, 1969, V16, P264, J ASSOC COMPUT MACH |
| 4 | 20 | MANI I, 1999, ADV AUTOMATIC TEXT S |
| 5 | 18 | RADEV DR, 2004, V40, P919, DOI, INFORM PROCESS MANAG |
| 6 | 15 | SALTON G, 1988, V24, P513, INFORMATION PROCESSI |
| 7 | 15 | SALTON G, 1997, V33, P193, INFORM PROCESS MANAG |
| 8 | 13 | PORTER MF, 1980, V14, P130, PROGRAM |
| 9 | 13 | ERKAN G, 2004, V22, P457, J ARTIF INTELL RES |
| 10 | 11 | LEE DD, 1999, V401, P788, NATURE |
| 11 | 11 | BARZILAY R, 1997, P10, P ACL WORKSH INT SCA |
| 12 | 11 | LIN CY, 2003, P71, P HLT NAACL |
| 13 | 11 | YEH JY, 2005, V41, P75, DOI, INFORM PROCESS MANAG |
| 14 | 11 | LEE DD, 2001, V13, P556, ADV NEUR IN |
| 15 | 10 | GOLDSTEIN J, 1999, P121, P 22 ANN INT ACM SIG |
| 16 | 10 | GONG Y, 2001, P19, P ACM SIGIR C R D IN |
| 17 | 10 | SALTON G, 1983, INTRO MODERN INFORM |
| 18 | 10 | LIN CY, 2004, P74, P WORKSH TEXT SUMM B |
| 19 | 10 | HEARST MA, 1997, V23, P33, COMPUT LINGUIST |
| 20 | 10 | OTTERBACHER J, 2005, P915, P HLT EMNLP |

Tabla 4: Autores más citados en el Dominio de la Sumarización

### 1.4.7.2.1.1- Colaboración entre autores

La gran producción autoral hace posible la colaboración entre autores y la creación de redes sociales de conocimiento, en esta figura pueden apreciarse como grandes centros de redes autores a Salton, Edmudson, Lunh, Radev, y Kupiec, autores que han sido centro del desarrollo de la minería de texto a nivel internacional (Ver Figura 1).

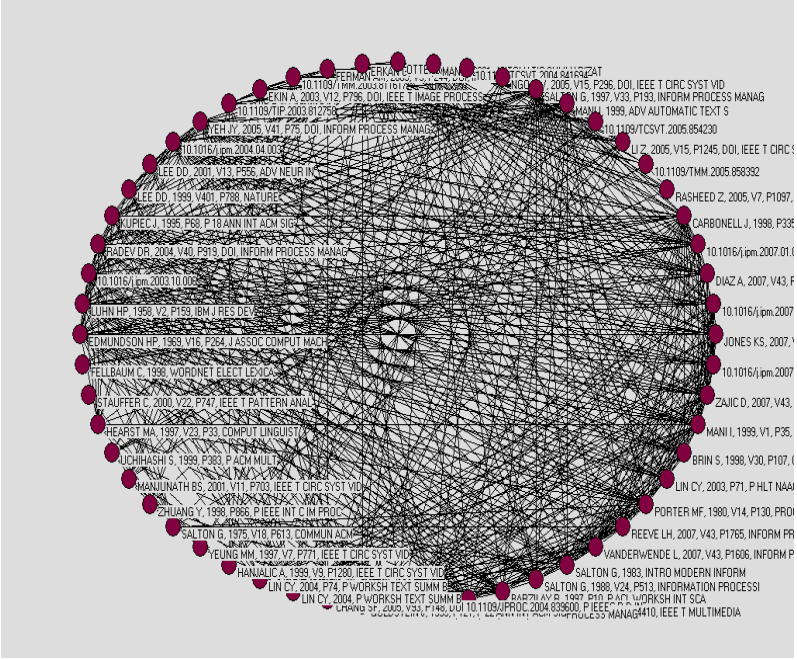



Figura 1: Relación entre autores en el tema Resumen automático

### 1.4.7.2.2.- Autores más productivos

La productividad autoral en esta temática puede dividirse en dos partes: autores corporativos y autores individuales. Dentro de los autores corporativos son indiscutibles los aportes de IEEE al desarrollo del tema, produciendo series monográficas, proceedings de eventos, etc. A esta entidad se deben la mayoría de los trabajos sobre sumarización al igual que a IEEE COMPUTER SOCIETY, quien dista mucho de la producción de IEEE, porque se ha estado especializando en los últimos años en artículos dedicados a aplicaciones y métodos de sumarización. Si de producción individual hablamos es importante hablar de los trabajos de DIVAKARAM, A y KACPRZYK, dedicados a métodos

de construcción de sumarios, algoritmos e implementaciones prácticas de diversos tipos de sumarios. Es importante destacar que no existen trabajos desarrollados por el gremio de la Ciencia de la Información entre los punteros, si no por autores de la Ciencias Computacionales, lo que demuestra que este dominio a pesar de ser un tema interdisciplinario ha sido muy poco estudiado por otras disciplinas (Ver Tabla 5).

| No | Autor | Cantidad | % |
|---|---|---|---|
| 1 | IEEE | 389 | 17.14% |
| 2 | IEEE COMPUTER SOCIETY | 35 | 15.42% |
| 3 | DIVAKARAN, A | 31 | 13.66% |
| 4 | KACPRZYK, J | 25 | 11.01% |
| 5 | ACM | 23 | 10.13% |
| 6 | FURUI, S | 17 | 0.74% |
| 7 | LI, WJ | 17 | 0.74% |
| 8 | LIU, Y | 17 | 0.74% |
| 9 | WILBIK, A | 17 | 0.74% |
| 10 | ZADROZNY, S | 17 | 0.74% |
| 11 | ACL | 15 | 0.66% |
| 12 | HE, YX | 15 | 0.66% |
| 13 | ISCA | 14 | 0.61% |
| 14 | MOUADDIB, N | 13 | 0.57% |
| 15 | PARK, S | 13 | 0.57% |
| 16 | PEKER, KA | 13 | 0.57% |
| 17 | RASCHIA, G | 13 | 0.57% |
| 18 | BOUGUILA, N | 12 | 0.52% |
| 19 | LI, Z | 12 | 0.52% |
| 20 | CHANG, SF | 11 | 0.48% |

Tabla 5: Autores más productivos en el campo resumen automático

### 1.4.7.2.3.- Países más productivos

El país más productivo en el tema del resumen automático es Estados Unidos de Norte América, de donde procede el 33.6 % de lo que se producen en el tema a nivel internacional. Le siguen en producción China, Japón Canadá y Taiwán, esto se corresponde totalmente

| No | País | Cantidad | % |
|---|---|---|---|
| 1 | USA | 763 | 33.6% |
| 2 | PEOPLES R CHINA | 382 | 16.8% |
| 3 | JAPAN | 171 | 7.53% |
| 4 | CANADA | 104 | 4.58% |
| 5 | TAIWAN | 88 | 3.87% |
| 6 | SOUTH KOREA | 77 | 3.39% |
| 7 | SINGAPORE | 63 | 2.77% |
| 8 | ENGLAND | 57 | 2.51% |
| 9 | FRANCE | 56 | 2.46% |
| 10 | GERMANY | 51 | 2.24% |
| 11 | SPAIN | 51 | 2.24% |
| 12 | ITALY | 47 | 2.07% |
| 13 | GREECE | 44 | 1.93% |
| 14 | INDIA | 42 | 1.85% |
| 15 | POLAND | 37 | 1.63% |
| 16 | BRAZIL | 35 | 1.54% |
| 17 | AUSTRALIA | 29 | 1.27% |
| 18 | CZECH REPUBLIC | 20 | 0.88% |
| 19 | AUSTRIA | 15 | 0.66% |
| 20 | NETHERLANDS | 15 | 0.66% |

Tabla 6: Países más productivos

### 1.4.7.2.4.- Años de mayor Publicación

A partir del año 92 la producción en este tema creció, sin embargo en el 2010, volvió a decrecer debido al estancamiento de las técnicas de tratamiento de texto, ya que son en su mayoría aplicaciones de corte matemático, alejadas de la lingüística y del análisis de dominio. El año de mayor publicación es el 2008, donde se publicaron 10 artículos más que en el 2007(Ver Tabla 7).

| No | Año de Publicación | Cantidad de Registros | % |
|---|---|---|---|
| 1 | 2008 | 340 | 14.9% |
| 2 | 2007 | 330 | 14.5% |
| 3 | 2006 | 278 | 12.2% |
| 4 | 2005 | 211 | 9.2% |
| 5 | 2004 | 187 | 8.2% |
| 6 | 2009 | 186 | 8.1% |

| | | | |
|---|---|---|---|
| 7 | 2003 | 159 | 7.0% |
| 8 | 2002 | 138 | 6.0% |
| 9 | 2000 | 87 | 3.8% |
| 10 | 2001 | 79 | 3.4% |
| 11 | 1998 | 53 | 2.3% |
| 13 | 1999 | 48 | 2.1% |
| 14 | 1997 | 37 | 16307% |
| 15 | 1991 | 18 | 0.79% |
| 16 | 1996 | 13 | 0.57% |
| 17 | 1994 | 12 | 0.52% |
| 18 | 1995 | 12 | 0.52% |
| 19 | 1992 | 10 | 0.44% |
| 20 | 2010 | 10 | 0.44% |

Tabla 7: Productividad por años en el tema resumen automático

### 1.4.7.2.4.- Editoriales más Productivas

No es una casualidad que el BIOMED CENTRAL LTD encabece el patrocinio de las publicaciones, debido a que la mayoría de las aportaciones prácticas y metodológicas en el terreno del resumen automático emanan de la actividad biomédica. También aparece como muy importantes los trabajos de ACADEMIC PRESS INC ELSEVIER SCIENCE, una editorial de fuerza en el tema que se trata. El resto de las editoriales producen entre 3 y 1 artículos.

| *CANTIDAD* | *Editoriales* |
|---|---|
| 5 | BIOMED CENTRAL LTD |
| 5 | ACADEMIC PRESS INC ELSEVIER SCIENCE |
| 4 | WORLD ACAD SCI, ENG & TECH-WASET |
| 4 | SCIENCE PRESS |
| 3 | ELSEVIER SCI LTD |
| 3 | WORLD SCIENTIFIC AND ENGINEERING ACAD AND SOC |
| 3 | WORLD ACAD UNION-WORLD ACAD PRESS |
| 3 | JOHN WILEY & SONS INC |
| 3 | INSTICC-INST SYST TECHNOLOGIES INFORMATION CONTROL & COMMUNICATION |
| 3 | ATLANTIS PRESS |
| 2 | SLOVAK UNIV TECH BRATISLAVA |
| 2 | NATL SUN YAT-SEN UNIV |

| | |
|---|---|
| 2 | IEICE-INST ELECTRONICS INFORMATION COMMUNICATIONS ENG |
| 2 | INT INFORMATION INST |
| 2 | SCIENCE CHINA PRESS |
| 2 | TAYLOR & FRANCIS INC |
| 2 | COLIPS PUBL |
| 2 | UNIV SZEGED, DEPT INFORMATICS |
| 2 | YELLOW RIVER CONSERVANCY PRESS |
| 2 | WORLD PUBLISHING CORPORATION |
| 2 | ELSEVIER SCIENCE INC |
| 2 | SPRINGER HEIDELBERG |
| 2 | ICIC INT |

Tabla 8: Revistas más productivas en el tema resumen automático

### 1.4.7.2.5.- Relación entre las Materias

Este tema tiene un componente de aplicación y metodológico que lo lleva a ser una temática multidisciplinar. Las temáticas y los campos que abarca el resumen automático pueden verse en la figura (Ver figura 2). Si se analiza la relación intermaterias es posible encontrar grandes grupos que se unen con materias específicas. Los grupos de materias más relacionados son los de Ciencia de la Computación y Sistemas de Información y el de Ciencia de la Computación y Ciencia de la Computación-Métodos Computacionales, en los mimos se enrolan materias tales como:

- CIENCIA DE LOS MATERIALES
- GEOCIENCIAS
- GEOGRAFÍA
- ÓPTICA
- INTELIGENCIA ARTIFICIAL
- CIBERNÉTICA
- BIBLIOTECOLOGÍA Y CIENCIA DE LA INFORMACIÓN
- INVESTIGACIÓN EDUCATIVA

- TECNOLOGÍA FOTOGRÁFICA
- MEDICINA
- FÍSICA APLICADA
- SOCIOLOGÍA
- LINGÜÍSTICA
- NEGOCIOS

Este comportamiento demuestra que existen muchas aplicaciones y usos del resumen automático en diversas ciencias.

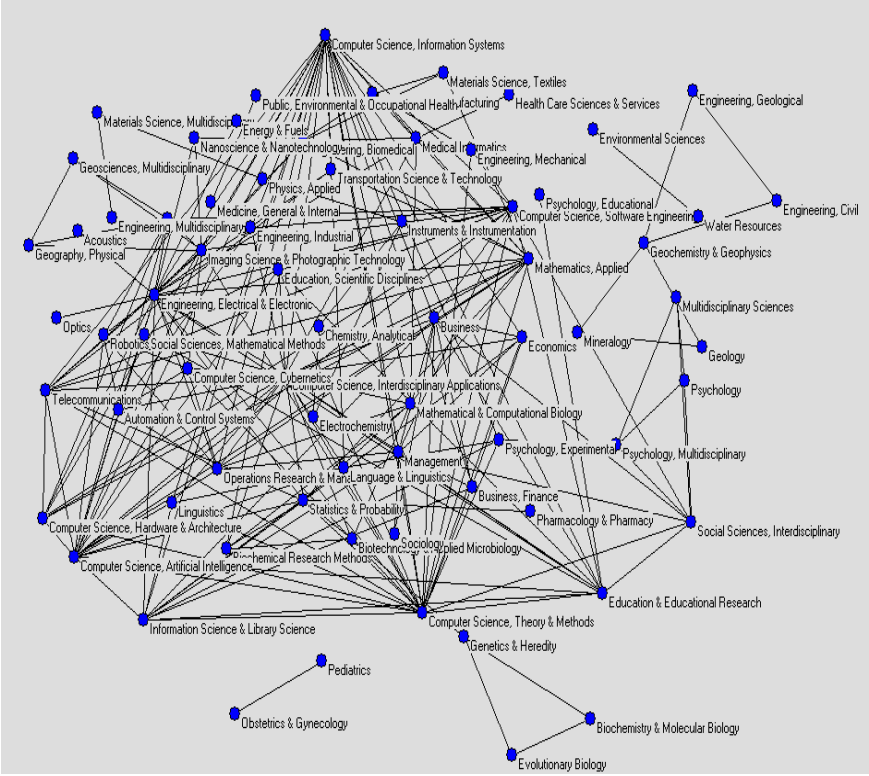



Figura 2: Relación disciplinar en el terreno resumen automático

### 1.4.7.2.5- Tipología Documental

Amplia y variada ha sido la tipología documental utilizada en la tesis, la mayoría en los documentos consultados fue ocupada por los artículos de revistas y los proceeding de eventos y conferencias, de obligada consulta por presentar las

propuestas más actualizadas en el tema. Se aprecia la revisión en la tesis de documentos no publicados, libros electrónicos, etc., lo que corrobora la gran cantidad de fuentes utilizadas en este estudio acorde con las exigencias investigativas del tema (Ver figura 3).

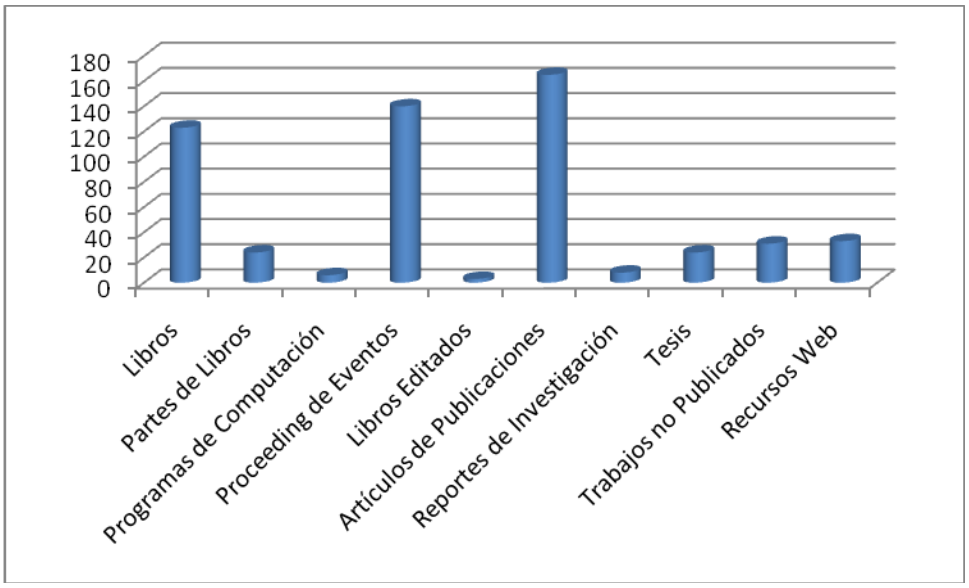



Figura 3 Tipología de Documentos utilizados en la investigación

### 1.4.7.2.6.- Autores más consultados

Los autores más consultados en la investigación han sido los que se destacan a continuación (Ver Tabla 9), el hecho de que sea José Antonio Senso el más consultado obedece a que ha desarrollado trabajos de implementación, evaluación y lexicometría inherentes al terreno donde se centra la aplicación de modelo (ingeniería de Puertos y Costas). Trabajos también medulares han sido los de José Antonio Moreiro, María Pinto Molina, Indejeet Mani, D. Paice, Sandor Domínguez y G. Salton, autores que despliegan sus aportaciones en el terreno de la implementación, la metodología, y el desarrollo de sistemas de minería de texto. El resto de los autores del trabajo oscila entre 3 y 1 consultas.

| *Autores* | *Trabajos* |
|---|---|
| Senso, J | 6 |
| Domínguez, S | 5 |
| Lin, C | 5 |
| Mani, I | 5 |
| Moreiro, J | 5 |
| Paice, C | 5 |
| Pinto, M. | 5 |
| Domínguez, S. | 5 |
| Salton, G. | 4 |
| Resto | 3-1 |

Tabla 9 Autores más consultados en la tesis

### *1.5.- Aportes de la tesis*

La **novedad** de esta investigación, reside en que es el primer estudio de los resúmenes documentales en entornos digitales hecho en Cuba desde la perspectiva de la Ciencia de la Información y el único modelo desarrollado para una universidad en Cuba.

El **aporte teórico** del trabajo es que logra describir los referentes teóricos metodológicos de la producción de resúmenes automáticos, sus tipologías y los procedimientos que se usan para la confección de sistemas de resumen, además la propuesta metodológica teóricamente aporta nuevas visiones del temática del resumen automático visto desde la Ciencia de la Información con un enfoque multidisciplinar, encausado a lograr un paradigma socio-cognitivo de Representación Textual.

El **aporte práctico** de la tesis es que se les ofrece a los investigadores y profesores de la Ingeniería de Puertos y Costas una herramienta que puede ser implementada y mejorada para satisfacer las necesidades de diversos usuarios a partir del estudio de sus necesidades informativas.

### *1.6.- Limitaciones*

Asociadas al autor más que al tema. La postura de análisis mutidisciplinar que se exige para construir los referentes teóricos de esta investigación requiere de un entrenamiento adicional en diversas materias, pues es difícil observar un fenómeno desde afuera y escribir sobre él. Esta investigación obligó a desarrollar conocimientos asociados a diversos campos del saber para lograr que los referentes del modelo pudieran construirse en la práctica, esto demandó la elaboración y formulación de nuevos argumentos cada vez que aparecían nuevas visiones sobre el tema, lo que hizo que el autor tuviese que establecer nexos entre la teoría de diversas especialidades para dar una coherencia teórica al modelo con vistas a su implementación.


# Referencias Bibliográficas

### *1.8.- Referencias Bibliográficas*

ALONSO, M. 1958. Documentación. *Enciclopedia del idioma: diccionario histórico y moderno de la lengua española.* Madid: Aguilar.

ARCO, L. 2005. *Corpus Miner.* Tesis de Maestría, Universidad Central "Marta Abreu" de las Villas.

ARCO, L. 2008. *Agrupamiento basado en intermediación diferencial.* Tesis Doctoral, Universidad Central"Marta Abreu" de las Villas.

BAPTISTA, P., FERNÁNDEZ, C. & HERNÁNDEZ, R. 2005. *Metodología de la investigación,* La Habana, Pablo de la Torriente.

BÉCUE, M. 1997. *Análisis Estadístico de Textos, cuarto seminario de capacitación de docentes,* PRESTA, Universidad de Concepción de Chile y Universidad libre de Bruxelles. Belgique. .

BOLELLI, L., ERTEKIN, S., ZHOU, D. & GILES, C. L. 2007. A clustering method for web data with multi-type interrelated components. *In Proceedings of 16th international conference on World Wide Web.* ACM Press.

CALPE, E. (ed.) 2008. *Diccionario de la lengua española,* Madrid: Espasa-Calpe.

CUÉ, J. L. 1988. *Estadística,* La Habana, Pueblo y Educación.

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DOMÍGUEZ, S. 2009. Satcol 6 Herramienta Para El Minado De Corpus y Construcción De Índices Automáticos. *In:* BETA (ed.). Santa Clara: Universidad Central de las Villas, Departamento de Automática.

DOMÍGUEZ, S. 2011a. Calculuscopora. beta ed. Santa Clara, Universidad Central"Marta Abreu" de las Villas: Departamento de Ingeniería Automática.

DOMÍGUEZ, S. 2011b. FOXCORP. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de Ingeniería Automática.

DOMÍGUEZ, S. 2011c. Puertotex. Santa Clata: Universidad Central de las Villas, Departamento de Ingeniería en Control Automático.

DOMÍNGUEZ, S. 2010. PROTEX. beta ed. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de automática.

ENDRES-NIGGEMEYER, B. 2005. SimSum: an empirically founded simulation of summarizing *Information Processing and Management,* 36**,** 659-682.

ENDRES-NIGGEMEYER, B., MAIRE, E. Y SIGEL, A. 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing & Management,* 31**,** 631-674.

FLORIA, A. 2000. *Evaluación Heurística* [Online]. Available: http://www.entrelinea.com/usabilidad/inspeccion/Heur.htm [Accessed 26.noviembre 2010].

GUADARRAMA, P. 2009. *Dirección y asesoría de la investigación científica,* Bogotá, Editorial Magisterio.

HARTER, J. & BUSHA, L. 1990. *Metodología de la investigación en bibliotecología y Ciencia de la Información,* La Habana, Editorial Félix Varela.

HERNÁNDEZ, A. 2006. indización y Resumen. La Habana: Universidad de la Habana.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

HORACIO SALDAÑO, O. n.d. *Tesis de grado. Metodología de la investigación.*

KERLINGER, L. 2006. *Metodología de la investigación*, McGraw-Hill Interamericana.

LIN, C. & HOVY, E. 1998. Automatic Evaluation of summaries using n-gram co-ocurrence Statistic. *In Proceeding of HLTNAACL.* EE.UU.

LÓPEZ-HUERTAS, M. 2008. Organización y representación del conocimiento: curso de doctorado. La Habana: Universidad de la Habana.

LOPEZ YEPES, J. 2005. *Las tesis doctorales: producción, evaluación y defensa.* . Sevilla:, Fragua.

MANN, W. & THOMPSON, S. 1990. Rhetorical structure theory: a theory of text organization.

MARIMÓN CARRAZANA, J. A. 2005. *Aproximación al estudio del modelo como resultado científico,* Santa Clara, CENTRO DE ESTUDIOS PEDAGOGICOS.

MENESES PLACERES, G. 2010. *ALFINEV: Propuesta de un modelo para la evaluaciónde la alfabetización informacional en la Educación Superior en Cuba* PhD., Universidad de la Habana.

NIELSEN, J. 1994. Heuristic evaluation. *In:* NIELSEN, J. & MACK, R. (eds.) *Usability Inspection Methods.* New York, NY.: John Wiley & Sons,.

NIELSEN, J. 2002a. *How to Conduct a Heuristic Evaluation* [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_evaluation.html [Accessed 26. enero 2011].

NIELSEN, J. 2002b. *Ten Usability Heuristics* [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_list.html [Accessed 21.enero 2011].

NÚÑEZ, I. 2005. *AMIGA.* Tesis Doctoral, Universidad de la Habana.

RUBIO, M. J. & VARAS, J. 2004. *El análisis de la realidad en la intervención social : métodos y técnicas de investigación.,* Madrid, Editorial CSS.

SALTON, G., WONG, A. & YANG, C. 1975. A vector space model for automatic text retrieval. *Communications of the ACM,* 18**,** 613-620.

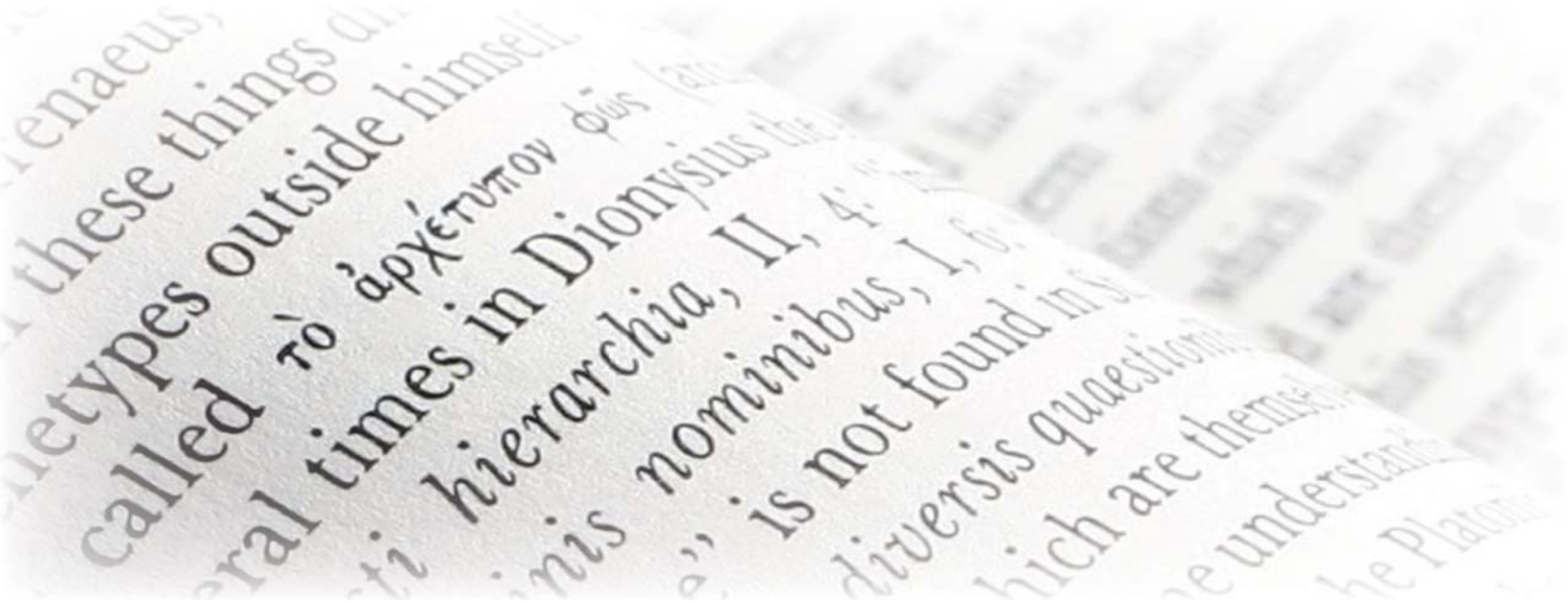
SATRIANO, C. & MOSCOLONI, N. 2000. Importancia del análisis textual como herramienta para el análisis del Discurso : aplicación en una investigación acerca de los abandonos del tratamiento en pacientes drogodependientes. *Cinta de Moebio***,** 1-23.

SENSO, J. & LEIVA, A. 2008. *Metamodelo para la extracción y desambiguación de textos científicos,* Santa Clara, Cuba, Universidad Central "Marta Abreu" de las Villas, Editorial Samuel Feijoó.

URÍAS, G. 2009. *Metodología de la Investigación,* Santa Clara, Universidad Central"Marta Abreu" de las Villas.

VAN DIJK, T. 2004. *Discurso y desigualdad,* Tenerife, Universidad de La Laguna.

ZALDUA, A. 2006. El análisis del discurso en la organización y representación de la información-conocimiento: elementos teóricos. *ACIMED* 14.




# Capítulo II

## El Resumen automático: fundamentos teóricos y metodológicos para su construcción

# Capítulo 2. El Resumen automático: fundamentos teóricos y metodológicos para su construcción

## *2.1.- Definición del concepto de resumen*

Muchas son las definiciones que existen sobre el concepto de resumen, sin embargo no todas se corresponden al desarrollo de esta actividad en los medios digitales. A continuación se mostrarán algunas de las definiciones más significativas sobre resumen. Para Alonso resumen es:"reducir a términos breves y precisos o considerar tan sólo y repetir abreviadamente lo esencial de un asunto o materia" (Dauden, 1982). Para Amat (Amat, 1988) el resumen es el "lenguaje libre a través de los indicadores, resumen descriptivo y resumen informativo". La profesora cubana María Josefa Daudén (Dauden, 1982) opina que un resumen es una breve exposición del contenido general de un documento (Dauden, 1982). Por su parte Lancaster (Lancaster, 1996) apunta las diferencias esenciales entre resumen y extracto y considera resumen sólo al trabajo realizado por un analista o especialista en información. Lasso (Lasso, 1969) expone sus criterios sobre resumen desde el punto de vista de la documentación y deja bien clara la función del resumen como medio de representación de los originales. En López Yépes (López, 1996) puede verse una definición muy ligada a la técnica propia de resumir y a sus entornos, sin embargo, la evolución de las técnicas sumistas hace que las definiciones evolucionen de acuerdo con el propio contexto donde se desarrollan los procesos extractivos. Los conceptos de resumen que nos brindan los autores anteriormente mencionados responden a los períodos en que formularon sus definiciones, que denotan al resumen como técnica y modo de representación textual. Un aspecto que no se aprecia en estas acepciones es el contexto automático, fenómeno surgido en la década de los 50 del siglo XX que parece ignorarse por la Academia.

Para definir el resumen en el entorno (digital) que nos atañe decidimos tomar a Pinto (Pinto, 2001) quien define el resumen de la forma siguiente: "resumen es

un nuevo documento representativo del original, que debe incluir todos los aspectos destacados del documento original siguiendo el estilo y la ordenación del documento original, y evitando cualquier apreciación y juicio crítico". Esta especialista española define la operación de resumir como un proceso que implica diversos cambios que deben experimentar los documentos textuales desde su estado inicial microestructural (o estructura de superficie léxico-sintáctica) hasta la obtención, y posterior descripción de su macroestructura (o estructura profunda lógico-semántica). La autora trata el resumen como una tarea de reducción informativa que, a la vez, constituye una operación de reconstrucción textual en modelo reducido. Además considera muy difícil cualquier intento de normalización de esta actividad.

Pinto (Pinto, 2001) propone un concepto de resumen que puede aplicarse al resumen automático, pues su definición abarca el conjunto de operaciones que distinguen a cualquier método de resumen. Esta autora también recoge las características de lo que debe ser un resumen de un documento electrónico, explicando las particularidades de un modelo ideal donde las extracciones se conviertan en verdaderas herramientas de mediación documental.

En opinión del autor el resumen automático debe ser una representación textual que mediante el uso de agentes inteligentes y el concurso de los sistemas de ontologías logre extraer de forma coherente los contenidos textuales de un documento estructurado en formato XML y, a su vez, pueda servir de intermediación hipertextual entre el usuario y el documento original amparado por las técnicas de minería de textos.

### 2.1.1.- Minería de Texto: herramienta multidisciplinar

El procesamiento automático de grandes cantidades de datos para hallar conocimiento, es la misión primaria del área de Descubrimiento de Conocimiento en Bases de Datos o KDD (*Knowledge Discovery from Data base*) definido por Lezcano (Lezcano, 2002) como un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia,

comprensibles a partir de datos, o como la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos.

Por su parte la minería de datos (*Data Mining*) es considerada por Senso y Eíto (Senso and Eíto, 2004) como una fase del KDD que integra los métodos de aprendizaje y estadísticas para obtener hipótesis de patrones y modelos. Indiscutiblemente esta se erige como una herramientas útil para realizar exploraciones profundas y extraer información nueva, útil y no trivial que se encuentra oculta en grandes volúmenes de datos estructurados (Senso and Eíto, 2004).

La diferencia sustancial que existe entre la minería de texto y la minería de datos reside en que la última procesa información estructurada y, sin embargo, aproximadamente un 80% de la información que se genera en la red de redes está almacenada en forma textual no estructurada, de ahí que se desarrollen actualmente técnicas de minería de textos (*Text Mining*), que facilitan el hallazgo de patrones interesantes y útiles en un corpus de información textual no estructurada (Drüsteler, 2002). Este proceder necesita del concurso de varias esferas del conocimiento entre las que se incluye la recuperación de información, el análisis de textos, la extracción de información, el agrupamiento, el resumen, la categorización, la clasificación, la visualización, la tecnología de bases de datos, el aprendizaje automático y la minería de datos (Dixon, 1997).

Estas disciplinas por sí solas no representan el paradigma de desarrollo de la Minería textual. Sin duda la lengua se ha convertido en un elemento clave de la llamada Sociedad de la Información (SI), de ahí que la lingüística, la organización de la información, la semiótica, la psicología, la documentación y otras ciencias sociales aporten elementos imprescindibles en el desarrollo del Minería Textual (Text Mining).

La minería de textos está reconocida como un proceder que permite el uso efectivo de datos textuales. Como resultado de las aplicaciones de la minería de texto, los textos no estructurados pueden ser analizados automáticamente

pudiéndose extraer los conceptos más importantes, etiquetar los documentos con términos claves, resumirlos y hacerlos disponibles en una forma normalizada y explotable (Molones, 2004).

La minería textual o (*text mining*), también conocida como minería de datos textuales o descubrimiento de conocimiento desde bases textuales, intenta acciones que le confieren cierta similitud con la minería de datos: identificar relaciones y modelos en la información no cuantitativa. Es decir, proveer una visión selectiva y perfeccionada de la información contenida en documentos, sacar consecuencias para la acción y detectar patrones no triviales e información sobre el conocimiento almacenado en las mismas.

La minería de textos es un proceder de alta complejidad, pues su fuente de análisis, los datos textuales, son inherentemente no estructurados y borrosos. La minería de textos es un campo multidisciplinario que utiliza el conocimiento expresado en bases netamente textuales para usarlo en procesos analíticos y predictivos (Dixon, 1997).

Aunque la concepción actual de la minería de textos posee un carácter estratégico, aún no ha tenido un desarrollo paralelo a las posibilidades que se brindan de almacenamiento de la información no estructurada disponible.

### *2.2- La Sociedad Digital*

El auge del resumen viene dado por el desarrollo de una sociedad digital donde toda la información que se consume es producida y difundida por medios electrónicos. Esta sociedad se caracteriza según Rosman, (Rosman, 1996) por:

Convivencia de lo impreso y electrónico.

Tránsito de la cultura acumulativa a la selectiva.

Cambios de estructura y formato que involucran cambios de contenido funcional (movilidad, dinamismo, manipulación, disponibilidad).

Consolidación del concepto de “documento” como objeto destinado a informar por medio de los datos que transporta.

Reafirmación de la información en el paradigma económico.

Aumento de la interacción hombre-máquina.

Aumento de la participación del usuario como ente activo del sistema.

Un elemento importante dentro de la sociedad digital es la Industria de los Contenidos la cual “asegura la concepción, producción, gestión y difusión de los recursos en consonancia con las necesidades de la sociedad de la información donde la cultura desempeña un papel socioeconómico significativo, y cuyos principales elementos son los servicios, productos, aplicaciones y programas. Basada en la materia gris, presenta una estructura más compleja y fluida que las industrias convencionales, una estructura que deja de masificarse para ramificarse según Hernández (Hernández, 2007).

Para Negroponte (Negroponte, 1995) la sociedad de las redes representa un paso de lo táctil a lo digital en un mundo altamente conectado, lo que pone de evidencia la existencia de una cultura digital.

Este conglomerado electrónico ancla sus acciones en diversos principios como:

Interactividad: Enlace físico entre personas e industrias basadas en la comunicación.

**Hipertextualidad:** Enlace de los contenidos e industrias del conocimiento. Forma de almacenamiento y representación de información digitalizada. La hipertextualidad potencia la intra e intertextualidad, difuminando fronteras entre autor y lector. Émulo de la organización asociativa de la mente.

**Conectividad:** Fomenta el enlace de mentes e industrias de redes, desarrolla nuevos hábitos cognitivos y formas de colaboración.

El agente de transmisión de la sociedad digital es el documento electrónico en todas sus variantes, incluso la hipermedial, las cual es conocida como una combinación de hardware, software y tecnologías de almacenamiento incorporadas para proveer un ambiente de información multisensorial y se caracteriza por los siguientes aspectos:

**Multicanaliad**: un soporte y varios canales de comunicación.

Interactividad: un usuario integrado en un entorno que tiene elementos de lo real que rompe la secuencia lineal de los mensajes.

**Multidireccionalidad:** Itinerarios por usuario.

**Multireferencialidad:** diversificación y multiplicación de las fuentes de información. Esta cualidad también establece la relación del texto con su intertexto y su contexto.

**Ergonomía:** Integración de la visualización en el contenido.

Estos documentos forman parte de los llamados Sistemas Multimedia.

### 2.2.1- Sistema Multimedia

Un sistema multimedia es la integración de los lenguajes icónico, textual y auditivo en un producto único que incluye todas las potencialidades reales y expresivas de los distintos códigos de comunicación como la imagen, el sonido y el texto. Los Sistemas Multimedia según Hernández (Hernández, 2006) se caracterizan por:

- **Interactividad** (en la búsqueda de información y toma de decisiones).
- **Ramificación** (estructura arbórea de datos. Capacidad de personalización de las respuestas).
- **Transparencia** (ergonomía para el uso y acceso al contenido de los mensajes).
- **Navegación** (consustancial a la interactividad. Los recursos conceptuales permiten orientar y visualizar mediante estructuras multidimensionales).

### 2.2.2.- Hipertexto, hipermedia y resúmenes

Se asume en este acápite una visión de las formas de los textos digitales por la necesidad de reconocer su importancia y preponderancia en los entornos

digitales y los nexos indisolubles que los vinculan con los procesos actuales y futuros de extracción textual.

El desarrollo de Internet y la proliferación de documentos digitales en un ambiente virtual traen nuevas concepciones al ámbito del resumen. La existencia de textos con cualidades de multimedia revoluciona las características de la documentación. El uso de elementos de sonido, imagen, etc., dan a los documentos una nueva perspectiva dentro de las estructuras reticulares. El término hipertexto ha venido designando, a lo largo del tiempo, tanto a una forma determinada de estructurar la información, como a una idea basada en la abstracción.

Nelson (Nelson., 1992) define a la documentación hipertextual como un concepto unificado de ideas y datos interconectados, y la forma en la que esas ideas y esos datos pueden editarse en una pantalla de ordenador. Según Lowein (Lowein, 2004) los documentos hipertextuales y el texto plano son formas que conviven también con documentos hipermediales, los cuales operan con significados importantes que no están basados solamente en su estructura textual, sino en el aprovechamiento de la estructura de los hiperenlaces. Los significados no son propuestos en los documentos hipertextuales, son el resultado de un lenguaje específico donde la estrategia de navegación y las formas de interacción son elementos que proponen una semiología particular en este medio (Solten, (2005)). La hipermedia tiene características semiológicas que constituyen su lenguaje de interacción para ello utiliza diversas metáforas visuales y referentes icónicos que sintetizan sus elementos informativos.

Para Chacón (Chacón, 2006) el resumen automático como medio de mediación documental tiene diversas funciones en los entornos cibernéticos entre ella están las siguientes:

1. Sintetizar los elementos fundamentales de los contenidos y eliminar los accesorios.

2. Permitir al internauta o usuario seleccionar la información que le interesa y decidir si le conviene leer determinados artículos e, incluso, proporcionarle información suficiente para evitarle su lectura.
3. Servir como documento para la indización y/o para la creación de ficheros de búsqueda.
4. Es un instrumento de difusión de las informaciones aparecidas en las estructuras reticulares.

Van Dijk (Van Dijk, 1995) apunta que el resumen es un ente necesariamente subjetivo porque presupone decisiones personales y profesionales sobre qué información es la más relevante o importante y qué categorías totalizadoras deben elegirse. El profesional de la información decide qué es lo esencial del documento y lo distingue de lo secundario y es precisamente esta decisión en la que intervienen elementos subjetivos -actitud frente al documento, cogniciones sociales, conocimientos previos, capacidad analítica y de síntesis- que influyen en la realización del resumen y que pueden predisponer al usuario en su forma de leer el documento primario, aumentando así las posibilidades de manipulación. De esta forma el profesional puede crear las condiciones de lectura o no de la página original de acuerdo a la forma en que se presenta. La manipulación se produce por subjetividad ya que el usuario consulta el medio documental si le interesa y luego decide si consulta o no el original.

El resumen del texto verbalizado de un documento digital puede afectar también la difusión de información, reduciendo la capacidad de recuperación del documento primario si se realiza apegado a criterios que no concuerden con los del usuario, o porque provoque sugerencias que impliquen la no recuperación de ciertos documentos, bien por carecer de relevancia o por pensar que pueda sustituir su lectura por la del resumen. La integración de todos los elementos distintivos de los textos verbales propicia reglas que no deben ser perdidas de vista a la hora de integrar métodos de construcción sumista para documentos en entornos digitales.

Las potencialidades de la hipermedia y el hipertexto como medio documental no han sido explotadas del todo (psicología, 2004), sólo el texto plano en forma digital es asumido con más frecuencia en las incursiones sumistas. En el entorno audiovisual en dependencia de la realidad textual se distinguen dos modelos de resumen: textual y audiovisual.

Al resumir un documento hipermedial, hipertextual o texto plano debe obtenerse otro documento textual (audiovisual o texto simple) compuesto por segmentos seleccionados. Esta recopilación de esta tipología documental puede ser textual (representación de la imagen mediante el texto) o audiovisual (superposición de imagen y texto, lo que facilita el proceso de representación e interpretación, o por separado, resumen icónico y textual). Esto significa que el documento puede representarse mediante códigos icónicos y textuales.

La presencia textual y la superposición de las imágenes deben facilitar enormemente los procesos de interpretación y la posterior representación de los documentos hipermediales e hipertextuales, lo que permite que tales formas documentales puedan verse representadas mediante resúmenes de rango. Así, el documento hipermedial e hipertextual se autocontextualiza mediante la superposición de imágenes, sonido y texto.

Las formas textuales que coexisten en la Web son poco explotadas y para el desarrollo de nuevas representaciones textuales demandan una nueva forma de expresión y de análisis, donde la cibersemiótica se inserte en los procesos de análisis. No se puede olvidar que la hipermedia y el hipertexto son formas textuales de valores superpuestos que demandan nuevas modelizaciones para ser resumidas coherentemente.

### *2.3.- Normalización y Calidad en el Resumen automático: Métodos de Evaluación*

La calidad de los resúmenes en ambientes digitales se torna algo compleja. El resumen en su doble función de proceso/servicio tiene complejidades que se dan por la propia naturaleza del soporte y los medios que lo distribuyen. Según

Rittberger (Rittberger, 1997) existen diversos sistemas de calidad para la evaluación de extractos artesanales, es decir resúmenes hechos de forma manual. Este mismo autor explica que el número de ponencias, artículos, proyectos, etc. elaborado sobre el tema de evaluación en ambientes electrónicos es cada vez mayor.

Los factores que inciden en la calidad de los resúmenes documentales en ambientes digitales son el contenido[5] y la eficacia de la representación[6]. La calidad aplicada al procesamiento de la información bibliográfica (texto escrito) no puede ser la misma que se use para los documentos electrónicos. Los valores de calidad de los documentos asentados en formatos tradicionales son fundamentalmente:

Factores contextuales o internos (texto, contexto, conocimiento base).

Funcionales externos (objetivos informativos).

Cada documento, esté en un ambiente hipermedial o no, necesita de una estrategia de análisis (Pinto, 2001) y eso impone la necesidad de una nueva taxonomía textual donde la distinción entre los documentos esté sustentada en alguna clave de organización como la retórica[7]y los elementos expositivos[8]. El proceso de Evaluación de los resúmenes automáticos aún necesita del concurso de elementos cognitivos y metodológicos que hagan de él un procedimiento único y transparente. Existen dos lecturas del fenómeno evaluación, por un lado la Ciencia de la Computación defiende la idea de que un resumen es correcto si el texto se parece matemáticamente al texto fuente y por otro la Ciencia de la Información sustenta que la comprensión del texto resumido debe ser el elemento esencial sobre el que verse la evaluación de un resumen.

No existe un consenso acerca del proceso de evaluación de los resúmenes automáticos dentro de la Ciencia de la Información. Para Borko y Bernier (Borko

[5] Concebido como la esencia o sustancia enunciativo-informativa de los documentos.
[6] Métricas de similitud.
[7] Narrativa con cierto argumento a desarrollar.
[8] Si el argumento es múltiple o difuso.

and Bernier, 1975) el resumen es un tipo de texto breve, preciso, claro, con atributos y estilos únicos. Con esta afirmación los autores comienzan a declarar algunas cuestiones relativas a la calidad de los textos resumidos. Lancaster (Lancaster, 1996, 1990) sustenta en dos características de la calidad del resumen enunciando algunos elementos de valor:

La aparición de los elementos esenciales del texto original.

La descripción sucinta de esos elementos esenciales sin ambigüedad y con precisión.

Bron y Day (Bron and Day, 1993) describen 5 elementos básicos para lograr un resumen de calidad:

Eliminación de la información trivial y redundante.

Generalización.

Integración.

Selección.

Construcción.

En el mundo digital la calidad de un resumen reside en tres aspectos básicos:

Personas :(resumidor, entidad cognitiva o automática).

Cosas: (documentos, cantidades, entidad física).

Métodos extractivos (software).

Por tanto la calidad descansa en la relación intrínseca entre el paradigma socio-cognitivo, la textualidad del documento y la herramienta que extrae los contenidos textuales.

### 2.3.1.- Atributos de presentación del resumen automático

Los atributos que representan un resumen automático no sólo están relacionados con las cualidades cognitivas del resumidor. Para Lancaster

(Lancaster, 1990) estos atributos se desglosan en: conocimiento del entorno, del dominio y uso adecuado de las herramientas.

Los indicadores que se proponen para evaluar un resumen son los siguientes:

**Precisión:** Nos da una idea de si el agente llega a representar con claridad el corpus textual, y si lo hace con la exhaustividad necesaria como para reconocer si los elementos descritos en el texto representan realmente un grupo textual.

**Exhaustividad:** Evalúa la capacidad del agente para reconocer dentro de un documento o un conjunto de ellos el tema tratado. La exhaustividad se debe analizar mediante métricas de similitud y aproximaciones superiores e inferiores.

**Legibilidad:** Según Tenopir y Jacsó (Tenopir and Jacsó, 2007) existen software que permiten llevar a cabo la evaluación otorgando una puntuación final entre 1 y 0. En cuanto a la Cohesión y Coherencia (elementos básicos que analizan la unión sintáctica de las oraciones del párrafo y su orden lógico), y teniendo en cuenta que para algunos estudiosos no es necesario el concurso del usuario final del sistema, también se pueden evaluar mediante herramientas diseñadas al efecto. Esto ha creado tendencias a la subvaloración de un fenómeno que desde el punto de vista lingüístico tiene sus reglas y análisis específicos, regidos en la actividad comunicacional humana.

2.3.1.1.- Algunos Métodos de Evaluación del Resumen Automático

#### *2.3.1.1.1.- Evaluación intrínseca*

Una de las tareas más complejas en la investigación sobre resumen automático, ha sido sin dudas, la formalización de los métodos de evaluación de los sistemas desarrollados. Son conocidas la opiniones de diversos investigadores entre los que se encuentra Pinto (Pinto, 2001), en las cuales se sostiene el criterio de que los métodos de evaluación del resumen, deben tener un marco de aplicación híbrida para que la posición del usuario pese en el desarrollo de la actividad evaluadora y esta no sea asunto sólo de algoritmos matemáticos. Para autores como D´cunha, (D'Cunha, 2006) las técnicas de evaluación de resumen han evolucionado demasiado para anclarse sólo en lecturas tradicionales. Si bien en

décadas posteriores las técnicas de evaluación textual eran básicamente manuales, es decir, que eran personas las que juzgaban la calidad de los resúmenes, con el desarrollo de la modelación matemática y la topología comenzaron a gestarse métodos de evaluación automáticos, debido a que los métodos manuales ofrecían un coste elevado, que podía ser mitigado con el uso de artes maquinales.

En los trabajos investigativos de Amigó (Amigó, 2005) se ofrece una clasificación de los métodos de evaluación de resumen automático, esta clasificación ha sido muy utilizada en la literatura, sin embargo, sus criterios no son construidos a partir de una visión excesivamente pragmática de ese proceso, pues intenta eliminar el papel del evaluador en todos los procesos evaluativos, ya sean automáticos o manuales. Según este autor los dos métodos esenciales para evaluar un resumen son los intrínsecos (que se basan exactamente en la forma estructural del texto) y los extrínsecos (que tienen en cuenta la función a la que se destina el texto).

Según D'Cunha (D'Cunha, 2006) los primeros a su vez se dividen en métodos de evaluación basados en la coherencia del resumen y métodos de evaluación basados en la información mantenida para el resumen (los cuales a su vez pueden tener en cuenta las fuentes originales o resúmenes modelo realizados de forma manual" (Figura 4). Este autor concuerda con que se clasifiquen las técnicas en dos esenciales, pero discrepa en la organización que propone el autor debido a que es una clasificación totalizadora de un proceso muy complejo, por ende utilizamos en esta tesis otra forma más general que incluye dentro de los métodos intrínsecos los métodos manuales con paneles de jueces.


Figura 4. Métodos de evaluación de resúmenes. D´cunha (2006)

**Evaluación intrínseca**

Los métodos intrínsecos centran su análisis directamente en el resumen. El referido estudio, según Baldwin (Baldwin, 2000) debe tener diversos criterios de calidad, como la corrección gramatical o la cohesión del texto que constituye el resumen, como de cobertura informativa. Esta forma de evaluación necesita del concurso de evaluadores humanos, pues juicios como la legibilidad y el grado de relevancia de la información seleccionada en la formulación del resumen, se erigen como patrones de un modelo netamente cognitivo que no puede ser constatado con los modelos algorítmicos actuales. Para lograr altos niveles de calidad en este tipo de inspección de texto se necesita, como principio fundamental, la especialización de la temática. Otro inconveniente asociado a la intervención de jueces en la evaluación es el posible desacuerdo que se pueda producir entre ellos. Un desacuerdo elevado puede dejar sin efecto los experimentos.

**Grado de concordancia entre jueces**

Posibilita establecer un análisis concordante entre jueces mediante una expresión matemática o métrica que ajuste los valores de las observaciones. En

esta forma de evaluación el proceder más utilizado es el índice de Kappa según Siegel y Castellán (Siegel and Castellan, 1988), medida empleada en sistemas de clasificación de texto en el terreno de la lingüística computacional (Carletta, 1996).

**Calidad del Resumen**

En este tipo de evaluación intervienen un grupo de jueces, los cuales leen los resúmenes y realizan una valoración de los textos. En Harman y Marcu (Harman and Marcu, 2001) se observa cómo se invita a los jueces a la valoración de tres aspectos de los textos: la corrección gramatical, la cohesión y la organización global de las ideas. El análisis de los resultados de la aplicación de esta técnica refleja diversos inconvenientes de acuerdo con los criterios seleccionados. En primer lugar, muchos de los jueces tienen dificultades para distinguir entre cohesión y organización. En segundo lugar, muchas veces aparece la falta de concreción en las posibles respuestas que se denotan del proceso.

**Cobertura de información**

Es la forma de evaluación más frecuente para medir la relevancia de la información. Su aplicación consiste en comparar el resumen generado automáticamente con uno confeccionado manualmente, el cual se considera base o ideal. La falta de concordancia descrita entre los humanos parece apuntar a la inexistencia de un extracto de este tipo. Por tanto, puede suceder que diversos resúmenes que no coincidan con el resumen patrón puedan ser relevantes desde el punto de la cobertura informativa.

Los métodos de evaluación extrínseca se fundamentan en la detección del efecto que tienen los resúmenes sobre la realización de alguna u otra tarea. En algunas investigaciones se ha evaluado la efectividad en la comprensión de la lectura, como por ejemplo en los trabajos de Morris, et. al.(Morris et al., 1992) Sin embargo, esta técnica no es muy popular entre los especialistas. La forma de evaluación más importante de este método es la llevada a cabo en la

evaluación del sistema TRIPSTER, que ha permitido obtener resultados medianamente aceptables según Arco (Arco, 2007).

Todas estas formas de evaluación de resúmenes tienen el inconveniente de ser procesos meramente cognitivos, muy costosos para la realidad de INTERNET, además los resultados tienden a ser imprecisos, pues los mecanismos de percepción de los evaluadores pueden ser diferentes a la hora de valorar un producto.

#### 2.3.1.1.2.- Evaluación Extrínseca

La idea de evaluar un resumen mediante evaluación extrínseca con el objeto de determinar el efecto sobre la síntesis de algunos textos está descrita en trabajos de D´cunha, (D'Cunha, 2006). El basamento práctico de este proceder se basa en el sustento teórico de que la precisión del tema y la pertinencia de evaluar los errores del texto fuente es atribuida esencialmente a los métodos de análisis matemático. La evaluación requiere que la fuente de documentos sea relativamente corta, ya que debe ser leída en una cantidad razonable de tiempo. Sin embargo, si los documentos son demasiado cortos, no hay necesidad de resumirlos. En esta vertiente se insertan los siguientes métodos: Evaluación automática, Métodos Semánticos y los Métodos de Superficie.

#### Evaluación Automática

La evaluación entre los resúmenes se realiza mejor por los seres humanos, pero también puede ser calculada automáticamente. Existe gran cantidad de medidas de evaluación textual en la literatura, las cuales se pueden utilizar, sobre la base de diversos estudios según (Lin, 2004). Los modelos de evaluación actúan generalmente sobre el estudio matemático de las oraciones tomando el texto fuente como referencia, realizando comparaciones sobre frases, segmentos textuales, etc. También puede calcularse la precisión, esta medida es aplicable esencialmente al sistema o software de resúmenes. Se desarrolla esta métrica de evaluación apelando a la forma en que se extraen las oraciones, que

indiscutiblemente no constituyen resúmenes, pues son prácticamente ilegibles (Marcu, 1999).

Sin embargo, tales medidas son insuficientemente sensibles para distinguir entre varios tipos de resumen, pues los resúmenes poseen muchas diferencias en su contenido y en su estilo.

Otra alternativa al uso de oraciones es denominada Sen-Rango frase, un resumen donde se especifica en términos de clasificación los méritos del texto resumido. Para ello se utiliza el ranking de la evaluación del texto resumido comparándolo contra el resumen imagen. Esto puede aplicarse con f-measure mediante variaciones en su cálculo. Este indicador puede entonces compararse con una correlación mea-sure. Sin embargo, este proceder es mejor cuando se aplica en la extracción de frases textuales, lo que facilita un equiparamiento con las formas de evaluación humana. F-measure es aplicable para evaluar extractos, no resúmenes. La Utilidad de las medidas basadas en f-measure es que se basan en la selección de rasgos más específicos para juzgar el resumen, dando mérito a las sentencias y oraciones. La ventaja aquí radica en la medición de la informatividad, sin embargo, varían en su capacidad para discriminar elementos de presentación textual (Anexo 24).

Otro elemento para evaluar los resúmenes es descrito por Salton (Salton, 1997) analizando el contenido basado en las medidas de similitud léxica y el vocabulario, que en principio están orientados hacia ambos formas de representación textual: los extractos y los resúmenes. Una de las virtudes de esta técnica para evaluar contenidos es que el número de elementos de resumen (ngramas o bigramas) pueden, si se desea, ser ignorados en el cálculo de la similitud. Las desventajas de esta técnica es que estas medidas no discriminan muy bien entre los resúmenes que implican diferencias en el sentido expositivo del texto como por ejemplo: la negación, el orden de los vocablos, etc. Sin embargo, tales medidas pueden ser útiles para comparar entre los extractos, o para comparar resúmenes que se realizan bajo técnicas de corta y pega o sea sirven para confrontar dos resúmenes con el texto fuente o texto imagen.

También son útiles para la comparación de resúmenes más fragmentarios, como las listas de frases.

**Métodos Semánticos**

Esta forma de evaluación permite marcar el significado de cada frase en un texto, con una calificación subjetiva de la medida del resumen que abarca las proposiciones en la fuente (Van Slype, 1990). Sin embargo, el esfuerzo por lograr un resultado con esta técnica puede ser costoso, pues se requiere un sistema de anotación requerido y gran precisión en los métodos de clasificación implicados en la tarea. Un método más práctico de evaluación que el semántico es de estructura textual que se encuentra en el mismo enfoque de extracción de información (Paice and Jones, 1993).

**Métodos de Superficie**

En lugar de representar los conceptos en un nivel profundo, estos métodos permiten determinar si las ideas clave en la fuente (identificados por el subrayando de pasajes fuente, etc.) subyacen en el resumen. Una variante de este proceder se llevó a cabo en el método SUMMAC comparando el texto fuente y el resumen mediante preguntas y respuestas tarea Mani (Mani et al., 2008). Los pasajes de la fuente se caracterizaron mediante criterios de pertinencia a un tema.

El campo de la evaluación del resumen automático no es una parcela cerrada, la proliferación de métodos de evaluación es tal que aquí en esta investigación sólo se han descrito los esenciales. En los trabajos de Morris, et. al., (Morris et al., 1992), Maybury (Maybury, 1995), Mani (Mani et al., 1999), Saggion y Lapalme (Saggion and Lapalme, 2002), Baldwin (Baldwin et al., 2000) y Amigó (Amigó, 2005) se describen técnicas y procederes que versan sobre variantes que se sustentan en los principios que se han expuesto

### 2.3.2.- Normalización

Un fenómeno que es necesario describir en los contextos digitales es la normalización la cual ha sido vista siempre en los resúmenes clásicos, sin

embargo en los entornos digitales toma otras connotaciones ya que se complejiza mucho. La mayoría de los autores (ANSI, 1978, Borko and Chatman, 1963, Chaumier, 1986, ISO, 1976) coinciden en que en la normalización del resumen se dan los siguientes principios: Selección, Interpretación y Producción. Sólo Pinto (Pinto, 2001) resume y analiza las normas para la redacción, presentación y estilo de los resúmenes contenidas en diversos documentos generados por asociaciones y órganos editoriales como: American Bibliographical Center Servicies (ABD-Clio), Sociological Abstracts Inc., Chemical Abstracts Service (CAS), Biosciences Information Service (WTAS) de Reino Unido y la del Servicio de Documentación del Centro Latinoamericano de Administración para el Desarrollo (CLAD) de Venezuela, entre otros, y critica la norma internacional ISO 214-1976, así como las ANSI Z 39-14-1979 (Estados Unidos) y UNE (1990) sobre la elaboración de resúmenes.

En la crítica a estas estructuras normalizadas se observa un constante análisis de las perspectivas de construcción y presentación. La autora insiste en la poca consistencia y actualidad de los métodos. Describe la presencia de ambigüedad al referirse a los métodos de tratamiento textual. Insiste en la inobservancia de los nuevos métodos en la construcción de texto, ya que son muy pocas las especificaciones en lo referente a este apartado, pues la mayoría de ellas basa sus acciones en generalidades.

En este tema en el área nacional no existe más que la Norma Cubana (Normalización, 1983) para la Confección de resúmenes elaborada en los años ochenta, la cual no contempla los nuevos cambios en los procedimientos de construcción de resúmenes y su estructura. Esta metodología no trata en profundidad los procedimientos de confección de resúmenes documentales y es una copia fiel de la norma ISO, lo que le resta originalidad y adecuación a nuestro medio. En las Reglas de Construcción de resúmenes no se aprecian cuestiones referentes a la normalización de los resúmenes documentales en ambientes digitales, hecho que demuestra que pocas normas hacen referencia a este apartado, excepto las normas ANSI y UNE.

Otro aspecto que es necesario analizar en estos terrenos es el relacionado con la evaluación. El resumen es el producto de más complejidad de la etapa de Representación de la Información. Este producto en los ambientes digitales no sólo carece de normativas sino de mecanismos de evaluación. La Ciencia de la Computación ha elaborado medios para analizar la eficacia de los software que extraen los contenidos, pero dada la característica del resumen (procesos/servicio) se hace necesaria la evaluación de los mismos mediante la combinación de métodos matemáticos descritos por Mitre (Cremmins, 1985), Fukumoto (Fukumoto, 2003), Namba y Okumura (Nanba and Okumura, 2000)y Radev, D., Hovy, E. y McKeown, K. (Radev et al., 2002) y métodos cualitativos descritos en los trabajos de Pinto (Pinto, 2001). No se trata de borrar lo que siempre se ha usado como sistema de calidad, sino de establecer nuevos parámetros adecuados a l entorno donde se produce y se utiliza la información.

### *2.4.- Los Algoritmos de Agrupamiento: su función en la evaluación de Corpus Textuales y en la Selección de Términos*

Un terreno vacío en la Ciencia de la Información ha sido el estudio de los métodos de construcción y evaluación de términos y corpus textuales. Si bien se han detectado errores en la evaluación y la construcción algorítmica de los corpus y listas lexicográficas, muy poco se ha hecho en el estudio de los métodos de agrupamiento, siendo este el sustento esencial para la construcción de grupos de términos que en muchas ocasiones sirven de clave `para la selección de oraciones en determinados textos con estructura OMRC, una forma de construcción de resúmenes automáticos o simplemente en la selección de términos para la construcción de herramientas léxicas. En este acápite se intentará describir algunos métodos de evaluación de agrupamiento por su influencia tanto el desarrollo del resumen automático como en la definición de herramientas terminológicas, se aclara que sólo se presentan aquí los trabajos de evaluación de grupos textuales que tienen alguna relación con los métodos empleados en el trabajo.

El agrupamiento es un proceso subjetivo; el mismo conjunto de datos usualmente necesita ser agrupado de formas diferentes dependiendo de las aplicaciones" (Jain et al., 1999). Esta afirmación es totalmente cierta si se tiene en cuenta la posición del agrupamiento desde el punto de vista pragmático.

El proceder utilizado en la evaluación de los resultados de algoritmos de agrupamiento se denomina validación del agrupamiento según Theodoridis y Koutroubas (Theodoridis and Koutroubas, 1999) y Halkidi, Batistakis y Vazirgiannis (Halkidi et al., 2001), lo que constituye uno de los procesos más complejos dentro de la minería de textos. El término medida de validación de grupos se corresponde con un proceder matemático que propone una función que hace corresponder a un agrupamiento un número real, indicando en qué grado el agrupamiento es correcto o no (Höppner, 1999). Según Arco (Arco, 2008) estas medidas en su mayoría son heurísticas por naturaleza para conservar una complejidad computacional plausible.

### 2.4.1.- Clasificación de las medidas

Según Arco (Arco, 2008) las medidas de evaluación del agrupamiento se clasifican en: globales y locales, subjetivas y objetivas, internas, externas y relativas, y supervisadas y no supervisadas criterios que coinciden con los trabajos de Höppner (Höppner, 1999); Silberschatz y Tuzhilin (Silberschatz and Tuzhilin, 1996) y los de Kaufman (Kaufman and Rousseeuw, 1990).

Las medidas globales describen la calidad del resultado completo de un agrupamiento usando un único valor real, mientras que las locales evalúan cada grupo obtenido (Höppner, 1999). Las medidas objetivas miden propiedades estructurales de los resultados de los agrupamientos, por ejemplo, la separación entre los grupos y la compactación o densidad de los mismos (Halkidi et al., 2001). Tales propiedades no son un elemento cognitivo suficiente para mostrar al usuario, debido a que están desconectadas de la realidad contextual de las comunidades epistémicas, aunque su principal atractivo es que son independientes del dominio, según Silberschatz y Tuzhilin (1996). Las medidas

subjetivas evalúan teniendo en consideración la usabilidad de los grupos (Stein et al., 2000). Las investigaciones referidas a este tema han tenido menor preponderancia que las realizadas sobre medidas objetivas (Tuzhilin, 2002).

En diversos trabajos, se ha descrito la validación del agrupamiento en: medidas internas, externas y relativas entre ellos sobresalen los de Theodoridis y Koutroubas (Theodoridis and Koutroubas, 1999), Kaufman y Rousseeuw (Kaufman and Rousseeuw, 1990), formas de clasificación muy ligadas a la interpretación matemática de este fenómeno desde la Documentación. Estas últimas tienen un alto costo computacional según Halkidi (Halkidi et al., 2001). Las medidas internas evalúan el agrupamiento teniendo en cuenta los resultados del agrupamiento de términos o sea a partir de las cantidades que involucran los vectores de datos. Las medidas relativas se basan en la comparación del agrupamiento a evaluar con otros esquemas de agrupamiento o con resultados del mismo algoritmo con diferentes valores en los parámetros. Ejemplo si se quieren agrupar dos corpus lematizados habrá que analizar el proceso mediante comparación con otro tipo de algoritmo.

A continuación se mencionarán las principales medidas internas y externas reportadas en la literatura para la evaluación de particiones y cubrimientos, es importante destacar que las variantes para evaluar agrupamiento borroso no serán abordadas, por no ser utilizadas en esta tesis.

#### 2.4.1.1.- Principales medidas externas

Una medida externa es la entropía de Shannon (Shannon, 1948), la cual es una función de la distribución de las clases en los grupos resultantes. La entropía total para un conjunto de grupos es calculada como la suma de las entropías de cada grupo, ponderadas con el tamaño del grupo según Rosell, Kann y Litton (Rosell et al., 2004). En los trabajos de Steinbach, Karypis y Kumar (Steinbach et al., 2000) se usa la entropía como métrica de calidad, se obtienen mejores resultados cuando cada grupo contiene exactamente un elemento. Otras variantes de cálculo de la entropía por grupos y para el resultado del

agrupamiento en general se presentan en Karypis, Han y Kumar (Karypis et al., 1999) (Anexo 22).

Para efectuar el análisis de la calidad de grupos se ha apelado al uso de medidas externas basadas en medidas netamente subjetivas como: precisión y recall, procedentes del terreno de la recuperación de información y las adaptan a la validación del agrupamiento. Precisión (Pr) y Exhaustividad (Re) se calculan para un grupo j y una clase i dados, usando las expresiones Pr (i, j)=nij/nj y Re (i, j)=nij/ni, respectivamente. A partir de los valores de la precisión y exhaustividad se calcula la medida de evaluación F-measure, muy utilizada para evaluar agrupamiento y resumen automático. Según Arco (Arco, 2008) la efectividad de precisión y exhaustividad en el cálculo depende de un umbral α (0≤α≤1) (Frakes y Baeza-Yates, 1992) que debe ser determinado de antemano por el investigador. Esta medida dentro del resumen intenta determinar los elementos esenciales relativos a precisión y la cohesión del texto. Un valor global, Overall F-measure (OFM), se calcula usando el promedio ponderado de los valores máximos por clase de F-Measure sobre todos los grupos según Steinbach Karypis, y Kumar (Steinbach et al., 2000).

F-measure intenta capturar la similaridad que existe entre los grupos textuales del agrupamiento obtenido con los grupos de referencia o sea desarrolla un paralelismo entre el texto de partida y el texto resultante (Rosell, et al., 2004). En los trabajos de Larsen y Aone (Larsen and Aone, 1999) se desarrolla una versión de F-measure para un agrupamiento jerárquico, tomando por clase inicial el máximo valor de F-measure sobre todos los grupos incluyendo el máximo de niveles de la jerarquía posibles. Otras variantes de precisión y exhaustividad, micro-average precisión y micro-average recall, son utilizadas para evaluar el agrupamiento (Niu et al., 2004), es recomendable utilizar esta variante de evaluación si el texto que se quiere agrupar pertenece a un solo grupo textual y si el agrupamiento de referencia posee una única clasificación para cada objeto (Anexo 22).

Las medidas de evaluación anteriormente expresadas son aplicadas con regularidad en particiones textuales grandes. En Banerjee (Banerjee et al., 2005) se presentan variantes del cálculo de estas medidas sobre pares de puntos y frecuentemente se usan para evaluar cubrimientos. Los estudios de Rosell, Kann, y Litton (Rosell et al., 2004) se basan en el índice estadístico de Kappa que permite evaluar la relación existente entre dos textos relacionados con un texto fuente, considerando los dos textos obtenidos como intentos de agrupamiento textual a partir de un texto de referencia.

La similaridad entre dos grupos textuales es otra forma de evaluar el agrupamiento y constituye un parámetro que toma valores entre cero y el máximo valor de la entropía de los grupos, alcanzado cuando ambos grupos son idénticos según Xu (Xu, 2004). Es importante destacar que existen variantes normalizadas para este tipo de evaluación, pero no se han desarrollado expresiones matemáticas que permitan la valoración general del resultado obtenido.

En este tipo de evaluación de grupo la medida error del agrupamiento utiliza el número de asociaciones incorrectas o ausentes para medir la cercanía que existe entre el resultado del agrupamiento y la clasificación de referencia (Roussinov and Chen, 1999). Se considera, asociación en una partición a un par de segmentos textuales que pertenezcan al mismo texto, y asociación incorrecta a aquella que existe en la partición del texto de referencia y no en el resultado del agrupamiento, se denomina asociación ausente a aquella que existe en el resultado del agrupamiento y no en la partición de referencia.

Esta medida es más efectiva cuando se quiere evaluar particiones pequeñas, por tanto una variante normalizada en el intervalo [0, 1] se presenta para proveer una menor dependencia del tamaño de ambas particiones. A parir de estos postulados se desarrollan los trabajos de Roussinov y Chen (Roussinov and Chen, 1999) en los cuales se da una nueva definición de exhaustividad (cluster recall) y precisión (cluster precision) con el objeto de mostrar en qué medida el proceso de agrupamiento fue capaz de mostrar las asociaciones entre los textos

y cuál es la precisión de las asociaciones detectadas, respectivamente. Estas variantes de medición son muy utilizadas en algoritmos que obtienen cubrimientos (Anexo 22).

Otra medida para estudiar la calidad del agrupamiento textual ha sido la distancia Euclidiana (ver Anexo 3), la misma ha sido utilizada para medir la equivalencia estructural de dos grupos, siendo cero si los grupos son estructuralmente equivalentes y mayor si no lo son (Falkowski et al., 2006). Mediante el coeficiente de correlación de grupos es posible medir la equivalencia estructural de los textos mediante la división de la covarianza de la representación vectorial de los objetos por el producto de su desviación estándar. Este coeficiente, según Arco (Arco, 2008), toma valores entre -1 y +1 (los grupos son estructuralmente equivalentes).

Existen varias medidas basadas en la distribución de pares de objetos, entre ellas pueden contarse: el estadístico R; coeficiente Jaccard, índice Folkes y Mallows, estadísticos Huberts y normalizado (Kuncheva and Hadjitodorov, 2004). Estas medidas tienen un alto grado de complejidad y coste computacional y se aplican mediante las técnicas de Monte Carlo (Theodoridis and Koutroubas, 1999). Esta técnica calcula una función de densidad probabilística de los índices estadísticos definidos (Arco, 2008). La evaluación se realiza comparando el agrupamiento del texto con otros grupos textuales que pueden ser denominados grupos de referencia o con la matriz de proximidad.

#### 2.4.1.2.- Principales medidas internas

Los algoritmos de agrupamiento facilitan la evaluación del agrupamiento de cierto texto, pues a través de ellos pueden definirse diversas métricas para conocer variados aspectos de la estructura de un texto o de un agrupamiento como densidad (Brun et al., 2007). En la literatura se reportan investigaciones que se enrutan hacia el desarrollo de métricas que validan el agrupamiento de una manera no supervisada entre ellos puede verse: el índice Goodman-Kruskal que tiene una alta complejidad computacional (Goodman and Kruskal, 1954), el

índice C apropiado cuando los grupos tienen tamaños similares (Hubert and Schultz, 1976) también resultan tratados en la literatura los índices propuestos en Akaike (Akaike, 1974) y Schwartz (Schwartz, 1978), los mismos utilizan criterios de información, seguidos por los propuestos en Jain y Dubes (Jain and Dubes, 1988) y Bock (Bock, 1985) (Anexo 23).

El cálculo de la dispersión intragrupo y la separación entre los grupos textuales ha sido ampliamente trabajado, un ejemplo es el índice de Calinski y Arabas (Calinski and Arabas, 1974), utilizado recientemente en Maulik y Bandyopadhyay (Maulik and Bandyopadhyay, 2002) (Anexo 23).

Es posible también mediante métricas calcular la cohesión de los grupos. Se puede usar como una medida de validación de éstos Overall similarity, siendo este indicador una medida utilizada en la evaluación de la minería de textos para medir la cohesión a partir de la similitud de los pares de objetos en un grupo (Steinbach et al., 2000) (Anexo 23).

En la Ciencia de la Computación los procesos de evaluación tienden a ser geométricos. Un ejemplo son los índices de Dunn (Dunn, 1974) y sus generalizaciones (Bezdek and Pal, 1995) (Anexo 23). Los primeros pueden variar con relación a la medida de distancia entre grupos y el cálculo del diámetro del grupo que se utilice. El postulado original de Dunn se sustenta en el estudio del mínimo de todas las distancias entre pares de elementos para calcular la distancia entre los grupos, y considera el diámetro del grupo como la mayor distancia entre sus miembros. Esta posición teórica de Dunn permitió obtener grupos textuales muy compactos y muy bien separados, pero inconexos y desbalanzados. Bezdek en sus análisis de la teoría de Dunn sostiene que el índice Dunn es muy sensible al ruido (Bezdek y Pal, 1995).

Según (Arco, 2008) Bezdek elabora una variante en el cálculo de la distancia entre grupos mediante la estandarización, respecto al tamaño de los mismos y una nueva forma de cálculo del diámetro del grupo mediante el cálculo de la distancia de todos sus elementos al centro del grupo, también estandarizado por

su tamaño. Esta variante obtiene mejores resultados para diferentes dominios, pero es importante percibir que se hace referencia a un centro de grupo, y no todos los algoritmos trabajan con prototipos, ni la estructura de todos los datos son grupos con forma esférica.

Cinco generalizaciones de los índices de Dunn para validar grupos con diferentes formas hiperesféricas y disminuir su sensibilidad al ruido fueron propuestas en Bezdek y Pal (1998). El índice VSV es el indicador que calcula la suma pesada de las distancias entre grupos e intragrupos escogiendo la distancia mínima entre grupos y el promedio de las varianzas de los grupos como sus componentes, respectivamente (Kim and Park, 2001).

Otras propuestas de índices, en su mayoría con un alto costo computacional especialmente cuando el número de grupos y objetos es muy grande pueden verse en Xie y Beni (Xie and Beni, 1991), o en los trabajos que se han desarrollado por Dave (Dave, 1996); Milligan y Cooper (Milligan and Cooper, 1985). Una medida interna y global es la que describe el índice de separación de Poner (Höppner, 1999). En este punto de la investigación se han descrito los índices esenciales que permiten calcular la razón entre las distancias intragrupos y las distancias entre grupos textuales como son: índices Dunn, Davies-Bouldin, etc. Otros indicadores facilitan el cálculo de la suma pesada de esas dos distancias, por ejemplo SD, S_Dbw y vSV. En los trabajos de Kim y Ramakrishna (Kim and Ramakrishna, 2005) puede observarse un análisis del diseño y funcionamiento de estos índices, y las modificaciones propuestas en su estructura a partir de nuevas formas de cálculo de las distancias entre grupos e intragrupos para resolver las limitaciones encontradas.

En un trabajo de investigación desarrollado por Brun (Brun et al., 2007)se hace referencia al índice Silueta que es el promedio, sobre todos los grupos, del ancho de la silueta de sus puntos (Anexo 24). Según Arco, (Arco, 2008) dos cálculos fundamentales intervienen en la silueta de un punto: la distancia promedio entre el punto y todos los otros puntos en el grupo, y el mínimo de la

distancia promedio entre el punto y los puntos en otros grupos. Este índice tiene una alta complejidad.

Hasta aquí se ha intentado dar una panorámica de los elementos esenciales de la evaluación de resúmenes y el agrupamiento, de lo cual se entiende que se necesita mayor proximidad en el nivel teórico entre Ciencia de la Información y Computación, para resolver los postulados referentes a evaluación, que hoy toman posiciones totalmente diferentes, que a la vez pueden ser aportativas para el desarrollo de un modelo único de evaluación.

### *2.5.- Los agentes automáticos en el procesamiento del lenguaje natural*

Los agentes automáticos son los entes que intervienen en el proceso de desarrollo de los resúmenes automáticos. Los ordenadores van logrando espacios importantes en el desarrollo de las técnicas sumistas, ya que han sido capaces de sustentar los análisis basados en las técnicas de lectura y análisis/interpretación. El desarrollo de las técnicas lógico/sintácticas y lógico/esquemáticas han logrado darle al ordenador un lugar preponderante en el mundo de la representación textual. Las dificultades surgen en el momento en el que se pretende saltar los marcos del apartado semántico/ pragmático/ contextual de esas operaciones y, con ello, llevar a los sistemas a la realización de acciones hasta el momento vedadas para ellos (Pinto, 2001). Por esta razón se reconoce la incapacidad de los agentes para desarrollar y/o producir resúmenes coherentes. Según Pinto (2001) y Lancaster (Lancaster, 1990) los agentes son ineficientes a la hora de producir textos como unidades autónomas y coherentes.

Sin embargo, son capaces de producir otras formas textuales de menor categoría que son muy eficaces por su creciente flexibilidad e integración con los nuevos entornos donde el texto es cada vez más escalable. Para Pinto (2001) y Arco (Arco, 2007) el auge de los resúmenes documentales automáticos se da por las grandes posibilidades que brinda la sumarización como técnica de

extracción textual y la inserción de los productos resultantes del procesamiento del lenguaje natural en las redes de información.

Para resumir, mediante el uso de los robots se necesita la unión de los diversos paradigmas que formulen de forma coherente la posición de los aspectos sociocognitivos y situacionales del ser humano. En el resumen el agente automático necesita modelar a niveles pragmáticos la secuencia selección/interpretación y producción. (Véase figura 5).

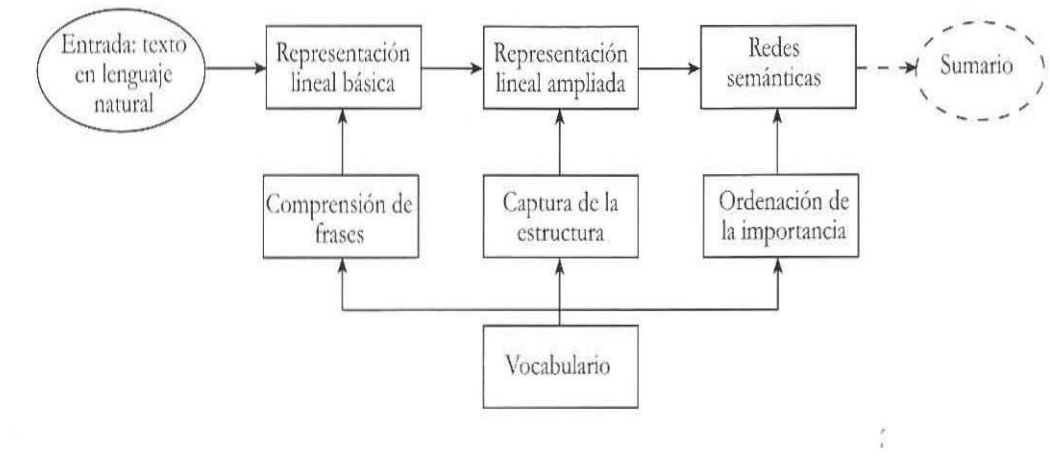



Figura 5. Secuencia de pasos del Resumen automático (Pinto, 2001)

### *2.6.- Paradigmas del resumen automático*

Los extractos generados por robots se insertan en un campo de actuación donde los paradigmas de acción definen su campo axiológico. Los extractos textuales automáticos, como todo producto de la Representación de la Información, poseen una connotación elevada, ya que al erigirse como producto documental complejo tienen la finalidad de: indicar, informar y sustituir al texto o los textos originales. En los entornos reticulares surgen diversas problemáticas que hacen que el usuario o receptor del producto quede parado ante el recurso como en un laberinto. Esta problemática hace que se necesite suministrar a los usuarios instrumentos de orientación adecuada. Los resúmenes en los espacios

electrónicos resuelven esta problemática mediante algunas cualidades intrínsecas como:

Etiquetar metadatos de documentos electrónicos.

Mediante la formulación de elementos de transmisión con el concurso de los hiperenlaces.

Para Pinto (2001) el ejercicio de resumen es una práctica documental que se inserta en el paradigma del procesamiento o Representación documental. Sus cualidades de mediación documental lo presentan como una herramienta de valor inestimable pues, según Lancaster (1990), sirve de mecanismo de información referencial para desarrollar funciones entrópicas semejantes al comportamiento humano. El resumen automático es una representación textual abreviada de los originales y se erige como un vehículo eficaz para la recuperación de la información en muchos sistemas automatizados. En algunos sistemas de bases de datos los extractos -gracias al agrupamiento textual- son utilizados como medio de indización de términos y por consiguiente como clave de búsqueda en texto libre (Pinto, 2001).

El resumen transforma las unidades discursivas, por tanto las problemáticas que debe sortear se dan en los terrenos de la comunicación y la enunciación. Como proceso científico el resumen es un proceso emergente anclado en diversos paradigmas y modelizaciones. El acto de resumir tiene un carácter multiparadigmático, pues sus basamentos se enfrentan al paradigma físico, cognitivo y lógico del procesamiento.

### 2.6.1.- Paradigma Físico

Aunque los estudios físicos o estadísticos de los textos nos brindan una lectura limitada y poco científica del texto (que necesita ser corregida por consideraciones de estructura y de significado) es innegable su campo de acción dentro de la lingüística computacional. La computación ligada al lenguaje le da gran valor al agrupamiento de los términos y su frecuencia de aparición con vistas describir el contenido documental (Medina and Pérez, 2007) . La

perspectiva estadística se ha usado como medio de evaluación y acarrea diversos problemas en el momento en que las palabras son repetidas, pues la escritura no está exenta de redundancias. El acto de comprender un texto es un proceso complejo mediante el cual los datos textuales son sintetizados. Gracias a las técnicas de análisis estadístico se agilizan los procesos extractivos y se utiliza menos memoria en los procesadores.

Existen otras variantes de análisis estadístico de los textos que son agrupados en el trabajo de Lin y Hovy (Lin and Hovy, 1998). Las matrices de términos documentales se erigen como conglomerados de términos contra términos, en las cuales las columnas representan el grupo de documentos que tiene un término en común. Estas agrupaciones de términos centran su base en la similitud y aproximación (Arco, 2007) y facilitan el establecimiento de las métricas de co-ocurrencia para facilitar el enlace entre las diversas representaciones de texto.

La contingencia (como también se le llama a esta forma de análisis en la cual se analiza la frecuencia de aparición de las formas textuales tanto a nivel monodocumental o multidocumental), según Pinto (2001), tiene el reto de desarrollar modelos donde en las evaluaciones exista una coincidencia con el espíritu del hablante.

### 2.6.2.- Paradigma Lógico

El desarrollo de los resúmenes está ligado al paradigma lógico, pues la lógica formal es la encargada de darle un sentido de comprensión a los fenómenos lingüísticos. La lógica se encarga esencialmente de los sistemas de inferencia que organizan los elementos del lenguaje en el acto de ser formulados. Setkin (Setkin, 2006) explica que la lógica formal es la entidad que estudia la forma de los enunciados, siendo esta la postura que más se aviene a nuestra investigación. La lógica formal analiza los términos tratados semánticamente bajo reglas de inferencia descritos por la Inteligencia Artificial. La introducción del formalismo como medio para el análisis de los signos de forma abstracta

tiene sus esencias en los procedimientos matemáticos descritos hace siglos. La manera en que se emplean y analizan los signos artificiales da al formalismo un papel preponderante en el análisis lógico de la simbología. El accionar del formalismo se aprecia en el tratamiento de los lenguajes documentales que poseen determinado léxico para lograr representar las cosas y los hechos.

La lógica formal es el único medio de formulación de un lenguaje extraído. Los lenguajes naturales mediante el concurso de la lógica formal deben convertirse en lenguajes artificiales, lo cual permitiría obtener productos analíticos rigurosos. Con esta afirmación el autor no quiere decir que sea necesario desarrollar una norma para convertir los textos extraídos en léxicos artificiales, sino que para la extracción textual la formalización es indispensable, de lo contrario persistirán los problemas de cohesión, coherencia y balance textual. En los trabajos de Pinto (Pinto, 2001) puede apreciarse cómo la representación informativa se compone de una serie de cuerpos relacionados que denotan su complejidad y los nexos de esta representación con otras formas representacionales. El formalismo proposicional posee algunas posibilidades de aplicación en los sistemas de resumen, ya que cualquier información debe estar sujeta a los procesos de memoria y del lenguaje (ver figura 6).


Figura 6. Secuencia Lógica del Resumen Automático

### 2.6.3.- Paradigma de la Psicología Cognitiva

Mediante este paradigma pueden conocerse los procesos sociales del conocimiento. Los diversos procesos cognitivos y las etapas cognitivas necesitan de una estructura de representación (redes semánticas y esquematizaciones). Los software para el procesamiento y tratamiento de la información no superan la calidad del ser humano como procesador de la información. Pinto (Pinto, 2001) enfatiza en que la práctica de identificar figurativamente a los ordenadores como procesadores de información sólo tiene el mérito de ser reconocida cuando se habla de temas informáticos. Esta afirmación es correcta, ya que son incapaces de desarrollar habilidades cognitivas como el hombre, como: imaginar, razonar, recordar, resolver problemas, intuir, conceptualizar y describir.

### 2.6.4.- Paradigma del Procesamiento de la Información

El paradigma del procesamiento de información evidencia las visiones que debe tener el procesamiento de la información dentro del resumen. Según Cornelius (Cornelius, 1997) este paradigma se centra en los puntos siguientes:

Diferentes sistemas de información como el hombre, las máquinas, los animales y las organizaciones procesan la información del mismo modo siendo crucial el software, es decir el modo algorítmico que desarrolla el procesamiento: Los sistemas de información en su generalidad tienden a trazar algoritmos o normas procesuales para la síntesis de la información. Las estructuras de representación tienden a formular los segmentos de análisis y las interconexiones iguales a los ordenadores y dejan sólo para las funciones analíticas un mínimo de habilidades cognitivas (Pinto, 2001). Se toma como modelo para los procesos cognitivos el pensamiento lógico consciente, no considerándose las fuentes del conocimiento intuitivas y emocionales. Se sigue construyendo herramientas basadas en algoritmos lógicos instalados como émulos de procesos donde los algoritmos no tienen igual formulación y desarrollo. Los algoritmos concebidos para representar información en los sistemas son estructuras cuyo valor es

eminentemente matemático y refleja una realidad sin variación, sin embargo el hombre desarrolla su base algorítmica en dependencia de diversos factores contextuales que pueden variar su sistema de análisis/percepción y modelación. Por otra parte, los sistemas poseen su paradigma donde el enfoque sistémico se erige como un ejemplo que facilita la concurrencia de algoritmos en todos sus procesos y entidades, lo que representa la obtención de las mismas sinergias, fruto de la repetición calcada de los mismos procesos, los cuales para reformularse o acoplarse necesitan de la mano inteligente de los recursos humanos de la Industria de la Lengua.

La comprensión es una teoría clásica dentro de la teoría de conjuntos: La comprensión dentro de la representación de la información está basada en los procedimientos y rutinas de la teoría de conjuntos. El acto de comprender tiene una base lógica y se edifica a partir de complejas formulaciones que son semejantes a los análisis descritos en la recuperación de la información.

Los procesos cognitivos pueden dividirse en partes de un proceso y constan de una serie lineal de elecciones: Cada proceso cognitivo es un ente independiente y a la vez es parte de otro proceso. Los procesos cognitivos son una cadena de sucesos que se ejecutan igual que la representación de la información, pero se diferencian de ella en el aspecto sociopsicológico, que si bien es diferente en cada individuo por ser una entidad individual y dinámica, es a la vez una proceso genérico e instrumentado cuando se habla de prácticas sociales igualitarias, como por ejemplo: comer, leer, escribir, resumir.

La percepción es una categoría clásica y denotativa: Mediante la percepción se denotan datos físicos, contextuales y connotativos, elementos claves para la representación de la información. El aprendizaje como teoría facilita el desarrollo y entrenamiento de los sujetos para desarrollar denotaciones e interpretaciones en el campo de la representación de la información.

El aprendizaje sucede de acuerdo con normas y principios, y consiste en la construcción de estructuras de conocimientos: En este sentido es indiscutible el

papel del aprendizaje para el uso de normas y acciones que toma la representación de la información para construir sus propias estructuras cognitivas.

Un sistema lingüístico es un mecanismo formal para la transmisión de información por medio de la manipulación de símbolos entre hombres máquinas y entre ambos. La representación de la información constituye un conjunto de herramientas que anclan sus acciones en los sistemas lingüísticos que interactúan con las máquinas y entre los individuos y estas. Esta proyección en la representación de los contenidos es hoy un paradigma y uno de los procesos emergentes de la Representación de la Información, la cual está demandando sistemas lingüísticos con suficiente carga semántica y pragmática capaces de transmitir, soportar y normalizar la transmisión de información. Estos sistemas adquieren una connotación diferente en el contexto digital pues las edificaciones textuales en estos entornos llevan una semiología particular, anclada en una postura semiológica denominada cibersemiótica donde el texto digital con todos sus atributos (textuales, icónicos y sonoros) se erige como herramienta de transmisión de un lenguaje particular.

La materia prima es cognitiva, y las emociones juegan un papel secundario: Los procesos cognitivos son importantes y primordiales para representar información. Los estados anímicos y otras situaciones culturales del dominio juegan un papel secundario, que influye de alguna medida en el procesamiento informativo, pues regulan las posturas que asumirá el representador de la información ante el objeto que desea representarse.

Hay una tendencia a ver la materia cognitiva como análoga del ordenador: Los ordenadores y los modelos teóricos sobre representación intentan ver los procesos cognitivos y los procesos algorítmicos como procesos de amplia diferenciación, sin embargo en la praxis estos procesos son iguales. Debido a esta visión igualitarista es que los procesos de representación automática son en múltiples ocasiones ineficientes y poco efectivos.

## *2.7.- El resumen como proceso y producto documental*

Muchas son las posiciones que asume el resumen dentro de los procesos de documentales y técnicos, estas posturas se sustentan en las técnicas de evaluación o estudio de los dominios, así como las técnicas de investigación social, utilizadas con gran profundidad en el trabajo con el resumen.

### 2.7.1.- Técnicas para el estudio de usuarios

Los humanos que utilizan los textos extraídos son importantes para el diseño y puesta en marcha de nuevos sistemas de resúmenes automáticos. Al ser el resumen un ente que tiene la cualidad de ser proceso/servicio se hace necesario entonces la detección de habilidades, hábitos informativos, necesidades, mecanismos de usabilidad Web, satisfacciones, expectativas, etc. Existen diversas técnicas que facilitan estos empeños entre las que se encuentra las técnicas cuantitativas y cualitativas.

#### 2.7.1.1.- Técnicas Cuantitativas

Las técnicas cuantitativas según Harter y Busha (Harter and Busha, 1990) facilitan el conteo y la cuantificación de la información en los estudios. Entre las que más se usan en el ámbito de modelaje de los sistemas de resumen se encuentra:

QFD (Quality Función Deployment) o Despliegue de la Función Calidad. (Pinto, 2001) opina que esta técnica que incorpora al diseño de resúmenes la voz del usuario haciendo que sus necesidades informativas sean monitoreadas de forma sistemática, convirtiendo al producto en una entidad generada a partir de necesidades. Esta técnica se ha visto en el sector de comercio y la industria. Los avances de este proceder en la edificación de resúmenes aún están en fase de estudio.

### 2.7.1.2.- Técnicas Cualitativas

Estas técnicas pueden ser directas e indirectas. Entre las indirectas[9] encontramos Métodos de análisis de contenido (textual, pragmático y discursivo): Ellas permiten reconocer el efecto de los productos y los procesos cognitivos sobre los documentos representados, es decir facilitan el reconocimiento de intenciones, ideas, creencias, circunstancias, situaciones estratégicas para lograr una representación semántica adecuada.

#### *2.7.1.2.1.- Técnicas Cualitativas directas:*

Método Delphi: Conjunto de cuestionarios estructurados que facilitan el desarrollo de discusiones abiertas entre especialistas lo que conduce al establecimiento de consensos y reduce las desviaciones. Puede ser aplicado a diversos contextos como: informacional, político, social, etc. Se perfila como un método confiable, pero altamente costoso.

El Método Escalar de Liskert: Basa sus acciones en las actitudes de los usuarios mediante la construcción de escalas.

## *2.8.- Procesos que intervienen en el trabajo con el resumen automático*

Para trabajar con un resumen documental se describen varios procesos que son necesarios para el desarrollo de herramientas coherentes y productivas. Los procesos que deben distinguir la actividad del resumen automático son los siguientes: Procesos de percepción y Selección, Procesos de Análisis\ Interpretación, Procesos de Síntesis /Producción y Procesos Complementarios Alternativos.

### 2.8.1.- Procesos de Percepción Selección

Es un proceso de lectura e interpretación de los datos textuales del texto. Mediante este proceso la base perceptiva lee e identifica elementos textuales

[9] Aportan información sobre las estructuras sociocognitivas y pragmáticas empleadas para representación y recuperación de la información.

que se corresponden con las características del usuario. En este proceso se dan dos procedimientos de gran importancia como la percepción de los datos por parte del agente y por el usuario final, que obtendrá el corpus textual transformado. Otro elemento dentro este proceso es la selección sintáctica que es capaz de describir las unidades léxicas y sintácticas dentro de la estructura textual. Este proceso se describe en Pinto (2001) y D´Cunha, (2006).

### 2.8.2.- Procesos de Análisis Interpretación

Describe los procedimientos utilizados en la lógica formal y su apego a esquemas de interpretación declarados a través de inferencias y casos situados con técnicas de procesamiento de algoritmos. La función lógico esquemática obliga al aparejo de los métodos interpretativos a la lógica formal en la que se establecen esquemas de análisis previamente concebidos. La actividad de interpretación también se apega a la semántica. La acción de los significados en los niveles interpretativos es de vital importancia para la comprensión de los textos que genera el ordenador, además es útil para lograr la comunicación con el usuario final. La parte pragmática de esta actividad está en lograr el desarrollo de los procesos de análisis e interpretación de acuerdo a las condiciones reales o sea de acuerdo a los escenarios y a las personas que usan la información extraída.

### 2.8.3.- Procesos de Síntesis Producción

Para Pinto (2001) muchos especialistas describen a este proceso como el mero acto de resumir. Este es el proceso donde se compone la información extraída después del análisis. En el caso de los agentes nos encontramos en la encrucijada de la poca actuación de las cualidades que caracterizan al autor. Estos asuntos en los agentes tienen otras características que se corresponden con las actuaciones de los procesos inferenciales (Namba, 1999). Según Van Dijk (Van Dijk, 1978) sin los procesos productivos y constructivos no es posible la expansión de cualquier texto sujeto a reglas específicas, lo que trae el

desarrollo de la escritura del texto de acuerdo con su estructura, estilo, requisitos y extensión.

Pinto (2001) opina que las relaciones entre las formas elocutivas y el contenido de las proposiciones, así como los modelos de linearización y los conectores lingüísticos forman parte de un sistema y éste puede usarse como medio de representación de las instrucciones de la producción textual. Todavía se necesita crear una fórmula explícita de las relaciones anteriormente descritas, lo que permitirá la construcción de un modelo informático que pueda aplicarse para la investigación y el apoyo de la actividad de la producción de textos.

### *2.9.- Desarrollo de las técnicas de extracción de texto verbalizado*

Los sistemas que se mostrarán en este apartado sólo se utilizan para la confección de resúmenes a partir de textos verbales, donde la palabra es el signo fundamental. Son viables para representar cualquier tipo de recurso ya que lo esencial en estos procedimientos no es el formato de salida, sino el elemento que se analiza.

Los sistemas automáticos están estructurados en consonancia con los procedimientos que se corresponden con los avances de las Ciencias de la Computación y dentro de ellas la Inteligencia Artificial. El primero en modelar estas actividades fue Salton quien utilizó una propuesta metodológica basada en el tratamiento textual con fórmulas de vectores.

Existen varios servicios que operan con sistemas de corte lógico, es decir mediante el uso de la inteligencia artificial y la minería de texto. Estas ofertas han logrado desarrollar al menos resúmenes informativos, ya que los volúmenes textuales dificultan los procesos extractivos, lo que provoca serias dificultades en la calidad de los extractos. Las dificultades que se aprecian en la calidad de los resúmenes se dan ante la imposibilidad de generar extractos para más de un dominio y en el trabajo con las diversas categorías de resumen, pues los estilos de construcción no se adaptan a la diversidad de los grupos. Por esto puede afirmarse que los métodos aún necesitan un perfeccionamiento que se adapte a

las características de Internet y a las de los diversos dominios que están en comunicación con la Red.

La posición que necesariamente tiene que tomar la ciencia ante este fenómeno es la confección de híbridos creados por ordenadores y hombres. Los sistemas lógicos no pueden suplantar la posición del ser humano y aunque ya existen redes asociativas con comportamientos complejos, lo cierto es que los problemas de construcción siguen sustentándose en los sistemas que operan con inteligencia artificial.

Han existido herramientas como SMART, ANES, pero ninguna ha logrado el desarrollo de resúmenes de calidad, pues pierden de vista elementos como la cohesión, la coherencia y el discurso propio de los dominios con los que interactúan. La posición más clara está en la existencia de un híbrido para realizar la labor de resumen en los medios digitales.

Los primeros métodos de resumen se centran básicamente en las investigaciones de Lunh y Edmunson (Maña, 2003) quienes en los años cincuentas y sesentas del siglo XX desarrollaron métodos anclados en principios estadísticos y en técnicas de extracción ayudadas por diversos análisis lingüísticos. Las décadas posteriores fueron poco productivas en este aspecto. Sin embargo, los años 90 del propio siglo XX patentizan la explosión de la producción científica en materia de métodos extractivos. A partir del modelo basado en el análisis espacio vectorial diseñado por Salton (Salton et al., 1975) se desarrollaron técnicas de construcción de texto de alta calidad.

Una evidencia del desarrollo de esta temática a escala universal son los libros editados por figuras como: Mani, Endres – Niggemeyer y Marcu,[10] así como los trabajos desarrollados en el campo de la Documentación por Borko (Borko and Bernier, 1975), (Pinto, 2001) y Lancaster (1990).

Como evidencia del desarrollo de la investigación en esta temática pueden tomarse también las secciones que al respecto aparecen dedicadas en diversas

[10] Grupo de autores que más publica en el tema del Resumen Documental.

publicaciones como: Computacional Linguistic, Information Processing Management, Journal of Information Science[11] así como los diversos talleres impartidos en disímiles conferencias a nivel Mundial. Con el Surgimiento de Internet se crea un nuevo escenario para el uso de los métodos de extracción. La industria de los contenidos comienza a lucrar con el desarrollo de herramientas para diversos fines. Surgen así los Sistemas Automatizados de Extracción de Noticias (ANES), los cuales a pesar de no resolver el problema de cohesión, coherencia y balance textual abren camino a una nueva dimensión en los Sistema basados en Inteligencia Artificial encargados de extraer los textos mediante arquitecturas de significado.

Los ANES son de corte extractivo y se apoyan en la estructura sintáctica del texto, lo cual permite extraer frases y oraciones, sin embargo este proceder acarrea problemas de cohesión y coherencia que generalmente se resuelven con un Sistema de Revisión de Textos. La ventaja de esta vertiente extractiva es que permite trabajar en entornos generales, pues no depende del domino, ni del volumen documental, ni de la tipologías textuales para realizar extractos.

El auge de la documentación digital en Internet trae como consecuencia el desarrollo de resumidores como Inxight (Xerox) y Copernic, además del fomento de sistemas en vías de desarrollo que operan online, como: SweSum y Extractor, que ofrecen resúmenes del español. Entre las investigaciones más actuales se aprecian como muy relevantes las desarrolladas por la Universidad de Ottawa y los proyectos del investigador de la Universidad de Columbia, Dragomir Radev. En el caso de España los esfuerzos se encaminan hacia la construcción de un prototipo de sistema resumidor de noticias periodísticas en el marco del proyecto Hermes.

Los Software de mayor calidad están sustentados por proyectos de investigación patrocinados por el Departamento de Defensa de los EE.UU. Entre ellos tenemos: DARPA y TIDES para la detección, extracción y resumen de

[11] Publicaciones Importantes en el campo de la Actividad del Resumen.

información multilingüe y el congreso de comprensión de documentos, Document Understanting Conference que propone un conjunto de investigaciones a gran escala con el fin de evaluar las técnicas y sistemas de resumen y comprensión de texto. Los sistemas anteriormente descritos utilizan la abstracción como medio de análisis, pues facilita el estudio del significado a nivel de frase. Esto permite una representación semántica del contenido del texto que puede utilizarse de conjunto con procedimientos de generación de lenguaje natural para confeccionar extractos. Las diversas etapas que abarca esta metodología obligan a establecer con claridad los dominios y sus características, por tanto sólo puede ser aplicada en contextos específicos donde se conozcan con exactitud los usuarios.

### 2.9.1 Tipologías de resúmenes automáticos

Un problema de la minería de textos, en la actualidad, es encontrar documentos relevantes sobre la información que se necesita, pero realmente la situación es incluso más compleja. Después de encontrar algunos documentos relevantes, el problema es encontrar el tiempo necesario para leerlos, por tanto, resumirlos puede contribuir a realizar una revisión en tiempo de sus aspectos relevantes.

Resumir textos automáticamente consiste en tomar una fuente de información, extraer contenido de ésta, representando los conceptos más importantes de cada documento y presentar el contenido más importante al usuario en una forma compacta y sensible para satisfacer las necesidades de la aplicación o del usuario, según (Mani et al., 1999).

Es un proceso que toma un documento como entrada, ofreciendo como salida un documento más corto, llamado documento sustituto, que contiene el contenido más importante del inicial. La importancia puede ser considerada a partir de diferentes puntos de vista: fragmentos asociados a las palabras clave, requerimientos de usuarios y relevancia a partir de tópicos seleccionados, entre otros, señalan (Jackson and Moulinier, 2002).

Existen dos tipos principales de resúmenes: los resúmenes abstractos (abstract) y los resúmenes extractos (*summarization*). Sus aplicaciones y formas de obtención difieren sustancialmente:

- Un extracto es un resumen que es construido, sobre todo, escogiendo los fragmentos más relevantes del texto, quizás con algunas pequeñas revisiones.
- Un abstracto es un resumen que describe el contenido de un documento sin presentar, necesariamente, algunas de las partes del contenido del texto original explícitamente.

En ambos casos, se puede pensar en el resumen como la compresión o la compactación de un documento. Un extracto realiza la compresión descartando el material menos relevante, mientras que un abstracto realiza la compresión de un modo más sofisticado, suprimiendo detalles y sustituyendo datos específicos con generalidades. Obviamente, se pudieran mezclar estos dos modos de compresión para obtener un resumen de mayor calidad como afirman Jackson y Moulinier, (2002).

Otra forma de clasificar los resúmenes es en genéricos (generis sumaries) y teniendo en cuenta las preguntas de los usuarios (Query-relevant summaries).

Los primeros se generan teniendo en cuenta todo el contenido del documento y los segundos resumen el contenido relevante a partir del entorno de la consulta realizada por el usuario.

La tarea de resumir textos puede ser dividida en dos fases:

a) construcción de una representación del texto.

b) generación del resumen, el cual puede incluir extracción de oraciones y/o construcción de nuevas oraciones.

La construcción automática de nuevas oraciones es una tarea extremadamente difícil y la mayoría de los sistemas generan un resumen extrayendo las

oraciones más relevantes del documento original, afirman Larocca, y Santos (Larocca and Santos, 2000).

Un resumen se puede realizar a partir de un único documento o de múltiples documentos relacionados o no. Este es otro aspecto que permite la clasificación de los resúmenes y da como resultado que existan resúmenes únicos (*single summarization*) a partir de la generación de textos pequeños y concisos que representan las ideas generales presentadas en un texto, o resúmenes que son extractos de múltiples documentos (*múltiple summarization*), según Larocca y Santos (2000).

Se han observado en diversos trabajos de doctorado como los de Maña (2003) y Arco (2007) (ajenos al campo de la Ciencia de la Información) el uso de la tipología indicativo para referirse a extractos generados de forma automática, cuestión que desde el punto de vista del estilo y la normalización es inconcebible, pues los agentes no pueden construir aseveraciones mentales al menos hasta el momento.

### *2.10.- Metodología del resumen automático*

Los resúmenes automáticos poseen una metodología particular que se aleja de los procesos clásicos reconocidos y experimentados en los textos tradicionales. Lo esencial en estos procederes es la ubicación de un método capaz para preparar resúmenes. Las máquinas computadoras son elementos imprescindibles para el desarrollo de los extractos automáticos. Sin embargo, los procesos de análisis y síntesis aún no permiten un salto cualitativo que facilite el desarrollo de la actividad sumista. Si bien los ordenadores son capaces de focalizar estructuras semánticas y lógico–sintácticas de gran complejidad la dificultad entre estas estructuras estriba en la desconexión de estas estructuras en los ambientes digitales (Pinto, 2001). El Procesamiento del Lenguaje Natural ha venido desarrollándose a partir de la década de los sesenta del siglo XX con herramientas creadas por Lunh, sin embargo fue MaThis y Rush (Mathis et al., 1973) quiénes describen con gran claridad los elementos básicos de un sistema

de resumen y su secuencia operativa. Mathis y Ruch, (1973) proponen la siguiente secuencia para el tratamiento automático de los textos:

- Lectura del documento original.
- Análisis aplicado a una serie de normas de selección y/o transformación.
- Construcción y composición del resumen.
- Edición e impresión.

La lectura se erige como un proceso complejo. Los sistemas de resumen tratan de emular con el resumidor humano, por tanto la lectura del documento desde el punto de vista de los agentes se complejiza, debido a que la misma es un proceso cognitivo. Generalmente estas herramientas operan con documentos usados frecuentemente por dominios determinados. La lectura desde un agente debe sortear diversos escollos que se intentan salvar, bien sea preeditado el texto (con la supresión de los elementos difíciles en el ordenador) o mediante la utilización de esquemas de codificación de los caracteres que no puedan ser leídos directamente. La etapa de más dificultad en las metodologías para la confección de resúmenes es la de la selección de los métodos de análisis y selección de normas. La mayoría de los sistemas cuyo accionar se fundamenta en procedimientos estadísticos que se reformulan constantemente según afirma Pinto (2001) son de corte estadístico.

Muchos autores como Pinto (2001), Arco (2007), Lancaster (Lancaster, 1993) y Hernández (Hernández, 2007) y analistas de las metodologías de tratamiento textual coinciden que los métodos sumistas que se utilizan en los entornos digitales sólo trabajan a nivel de frases textuales y que no hacen resúmenes reales debido a la falta de cohesión y balance del texto resultante. Uno de los problemas mayores en estos métodos es el paso de los extractos al nivel interpretativo por los defectos de precisión y ambigüedad que se dan en las diversas lenguas cuando de semántica se trata. Para mejorar esta situación se necesita desarrollar sistemas que reduzcan la semántica transformado los

corpus de texto a estructuras lógico-semánticas. Esta proyección sólo es asumida hasta el día de hoy por el ser humano. Uno de los avances que se da en aras de mejorar la problemática antes expuesta se denomina: Procesamiento del Lenguaje Natural.

Esta subdisciplina es un conglomerado de técnicas de inteligencia artificial que intenta diseñar sistemas que operen de como émulos del ser humano. Según Pinto (2001) los primeros estudios de Lenguaje Natural sobre agentes informáticos se describen en los años cuarenta del pasado siglo y la aparición de los ordenadores los convierten en un sector emergente de la lingüística computacional. En el Procesamiento del Lenguaje Natural subyacen teorías diversas como: Análisis de Discurso, Teorías Sociocognitivas, Modelos Sociopragmáticos de discurso, etc. Para Pinto (Pinto, 2001) en el PLN se ha generado un dominio con personalidad propia: el Knowledge Discovery (KD).

**Niveles del Procesamiento del Lenguaje Natural**

El lenguaje natural tiene niveles de procesamiento, los cuales facilitan la confección de propuestas metodológicas.

- **Nivel Morfológico:** Permite la construcción de analizadores morfológicos automáticos que recogen las principales ocurrencias de la unidades de la lengua. También en este nivel se dan estructuras gramaticales complementarias como: análisis de las formas flexivas, derivadas, y compuestas de las palabras. No logra eliminar la ambigüedad léxica, cuestión que lleva otros tratamientos luego que el extracto está listo.
- **Nivel Sintáctico:** Para Mann y Thompson (Thompson and Mann, 1988) este nivel permite estudiar las regularidades de las frases, propone fórmulas para eliminar la ambigüedad y es la base para desarrollar el enfoque proposicional que permite la interpretación. Desde el punto de vista lingüístico este nivel pretende enmendar los problemas que el nivel morfológico no resuelve, específicamente la ambigüedad terminológica. Este proceso funciona mediante el concurso de bases de conocimientos

que recogen todas las combinaciones gramaticales si/no aceptables entre términos (Pinto, 2001). Los analizadores sintácticos automáticos son capaces mediante estas bases de conocimiento de detectar o almacenar anáforas y catáforas. Conjuntamente con estas herramientas se hace necesario desarrollar el proceso steming que permite controlar la existencia de diversas palabras en distintas formas (plurales, tiempos verbales) reduciendo todas sus variantes a la fórmula canónica. Es importante destacar que el nivel sintáctico se vale de los parser que identifican los elementos de la frase y especifican sus relaciones, incluidos los marcadores discursivos y de cohesión.

- **Nivel Semántico:** Los analizadores semánticos permiten al ordenador el razonamiento de los significados, asignándole valor a las estructuras sintácticas descritas en el nivel anterior. En los Sistemas de Procesamiento de Lenguaje Natural el Nivel Semántico integra las llamadas Redes Semánticas[12]. La problemática de las redes semánticas se da fundamentalmente por la necesidad de bases de conocimiento sobre los términos y conceptos que utiliza cada segmento de usuarios. Esta situación se ha venido resolviendo mediante ontologías y otros elementos de la llamada cartografía documental.

- **Nivel Pragmático:** Para Pinto (2001) la comprensión de los diversos discursos no es sólo un conjunto de operaciones de codificación y decodificación de los mensajes simbólicos. El análisis morfológico debe estar integrado a procesos de comunicación, teniendo en cuenta no sólo los distintos conocimientos de los receptores sino la intencionalidad en el acto de la comunicación.

Los procesos sumistas no son procederes aislados, a pesar de tener metodología propia, ellos necesitan de procesos de organización y representación lingüística accesorios, es decir necesitan herramientas para

[12] Conceptos simples cuyas interacciones y combinatorias forman estructuras conceptuales complejas

describir y acceder a los resúmenes generados, de lo contrario se desarrollaría un sistema infoxicado por la falta de herramientas lingüísticas y representacionales. La Cibernética y las telecomunicaciones actualmente desarrollan herramientas que si bien no han solucionado del todo el problema, propician avances notables en el desarrollo de la indización y la búsqueda y recuperación de información.

### 2.10.1.- Cartografía Documental

El propio desarrollo de las nuevas tecnologías de información demanda nuevas representaciones más inteligentes y moldeables de caras a los ambientes virtuales. Esta aspiración se está viendo consolidada a través de las herramientas de análisis e inferencia que se construyen a raíz de los nuevos problemas de infoxicación con los que tenemos que convivir hoy en un mundo virtual. La presencia de cartogramas documentales ha dotado a la lingüística documental de herramientas que le facilitan el procesamiento cognitivo del lenguaje natural, entre ellos están: las ontologías, los mapas temáticos y las taxonomías.

#### 2.10.1.1.- Ontologías

Según definición de Tim Berners-Lee y Hendler: La Web Semántica es una extensión de la Web actual en la cual se dota a la información de significado bien definido para que tanto personas como ordenadores puedan trabajar cooperativamente (Barners-Lee et al., 2001). Es decir, utilizando la propia estructura tecnológica de la Web se pretende dotar a los recursos que en ella residen de algún elemento que añada información sobre la propia información, de manera que esta primera facilite la comunicación *“con sentido semántico”* entre las máquinas que forman la red y, por lo tanto, redunde en que las personas que la utilizan puedan obtener lo que desean con mayor rapidez y fiabilidad. Así pues, el objetivo último de esta nueva Web es que se produzca un intercambio de información efectivo y eficiente (Miller, 2001). Para que las máquinas puedan llevar a cabo esta función necesitan acceder a colecciones

estructuradas de información y a formalismos actualmente basados en la lógica matemática que les permitan tener cierto grado de razonamiento automático. Estas necesidades pueden cubrirse utilizando ontologías para anotar los recursos Web.

El término '*ontología'* (utilizado en filosofía para hablar acerca de una 'teoría sobre la existencia') ha sido adoptado por la comunidad de investigadores de inteligencia artificial para definir una categorización y las relaciones entre sus términos (Barners-Lee et al., 2001).

La categorización define clases y las relaciones entre ellas. Los elementos concretos son 'instancias' de esas clases, también conocidas como términos de la ontología. Como en cualquier clasificación, la relación básica entre términos es la de herencia, donde una clase *A* (subclase) es un tipo de la clase *B* (superclase), por lo que posee todas sus características.

Las relaciones semánticas que se definen entre las clases de una ontología pueden considerarse predicados que, junto a un conjunto de reglas de inferencia, permiten construir software (agentes) que realicen razonamiento automático. Cuando se define una relación de este tipo se puede especificar, entre otras muchas propiedades, el conjunto dominio de la relación y el conjunto imagen, es decir, las clases de las que parte la relación y las clases con las que se asocia según el significado de la misma. Por ejemplo, si se define una relación *está_compuesto* entre la clase *ordenador* y la clase *dispositivos_de_entrada*, y además existe una relación de herencia entre la clase *dispositivo_de_entrada* y la clase *teclado*, se podrá deducir que un teclado es una parte del ordenador; de forma que se sabrá que la instancia "teclado de conceptos" de la clase *teclado* es, además un dispositivo de entrada, una parte del ordenador.

Las ontologías se han visto como vocabularios que reflejan relaciones terminológicas que poseen reglas o normas para inferencias. Se han convertido en valiosas herramientas de cara a la indización de documentos digitales y el

desarrollo de la búsqueda y recuperación en la Web. Las ontologías son una agrupación de palabras o términos que describen un campo de saber completo, por tanto ver las ontologías como una posición independiente de los procesos extractivos es errónea. Son relaciones semánticas y desde la perspectiva de este trabajo se consideran como elementos de ellas a los facets maps, topic maps, y a las taxonomías herramientas de uso en las ontologías, pues son un modo de compartir conocimientos y sirven de comunicación entre especialistas de una misma rama en un sistema nodal. Para Senso (Senso, 2008) existen varias ontologías que pueden agruparse según sus tipos en cuatro categorías:

- Alto nivel: destinadas a describir todos los conceptos generales tales como el espacio, el tiempo, la materia, el objeto, el hecho, la acción, etc.
- De dominios: describen el vocabulario relacionado con un dominio genérico.
- De tareas: describen actividades, lo que puede resultar útil en las organizaciones.
- De aplicaciones: describen los conceptos conforme a un campo determinado o unas tareas concretas.

Son herramientas que operan a partir de conceptualizaciones que particularizan el léxico de una rama específica del conocimiento, por lo que están condicionadas a públicos muy específicos. Trascienden las acciones normativas de los sistemas de indización dando una apertura al mundo de la metainformación para el cual se convierten en meta-ontologías, permitiendo representaciones dinámicas y sobre un contexto previamente analizado. Facilitan la interacción a nivel de disciplina y subdisciplina en una materia determinada.

También proponen diversas relaciones que se observan en los lenguajes documentales que facilitan los accesos al conocimiento. Su única limitación es que las inferencias y las relaciones conceptuales deben estar en constante cambio con el desarrollo de la ciencia y el propio especialista debe realizar un

mapeo conceptual para el desarrollo de los mismos. Las ontologías y todas las estructuras o herramientas que corren sobre ellas proporcionan nuevas forma de representar y compartir el conocimiento gracias a la utilización de un lenguaje común previamente concebido para su aplicación. Poseen un formato flexible que aporta la usabilidad de los conocimientos representados abriendo paso así al intercambio de saberes, explicitan la condición del documento y sus relaciones, además facilitan los accesos múltiples y la reutilización de los conocimientos asociados.

Los facet maps, ontologías, y taxonomías constituyen recursos sobre los cuales se realizan repositorios para la organización del conocimiento abriendo puertas a la adquisición de información. Se erigen como instrumentos para la construcción de sistemas confiables sobre metadatos. Son utilizables para la normalización ya que reflejan casos en sus modos de construcción. Son Potenciales redes asociativas que aportan confianza y especificidad a las relaciones. Poseen un alto nivel de ergonomía, facilitan la posición ante el sistema de diversos usuarios. Conciben el tratamiento ponderado del conocimiento acumulado para recuperar información de forma automatizada ya que generan formas y procedimientos que facilitan la edificación de la estructura semántica de un sistema de Representación de Información. A la hora de diseñar una ontología debemos tener en cuenta 5 cuestiones clave: claridad, coherencia, extensibilidad, especificidad y precisión.

**Ontologías y Resumen**

El uso de ontologías en el proceso de summarización de textos ha sido uno de los nuevos procedimientos desarrollados en el resumen automático. Las técnicas de trabajo van desde el uso los clásicos lexicones computacionales hasta el resumen de videos.

Una de las primeras aproximaciones al trabajo con ontologías en el resumen es el trabajo teórico de Zang (Zang et al., 2010), en el que el autor hace un análisis de las posibilidades de la semántica como medio para hacer más escalables los

sistemas de resumen. Los autores parten de la idea que supone que la reducción de la cardinalidad hace más eficiente el escalado semántico. La noción de conocimiento semántico se basa en algunas técnicas de granularidad[13] computacional, summarización de textos y basamentos teóricos sustentados en la lingüística. Los principios en que se basan estos autores para construir su sustento operacional están en consonancia con lo siguiente:

- Principio de Máxima Cobertura
- Principio de Mínima Cobertura
- Principio de Variables de Ajuste

Otra dimensión del uso de las ontologías en la sumarización está en el trabajo de Zhang (Zhang et al., 2007) en el que se describe un método para la selección y comprensión de la lógica ontológica, para ello desarrolla una metodología capaz de analizar las sentencias declaradas en RDF , aprovechando su semántica para determinar las posibles construcciones de una nueva ontología. La salida de la consulta RDF se basa en un grafo de centralidad.

En esta misma línea de trabajo aparece el de Hu (Hu et al., 2004) donde se presenta un método para generar resúmenes de un solo documento teniendo en cuenta la máxima integridad y la mínima redundancia. Implementa un vector semántico basado en la representación de clases en diversas unidades lingüísticas de un documento a través de la ontología Hownet, capaz de mejorar la calidad de la representación, debido a que sus resultados de implementación a nivel de algoritmo superan los de otros modelos de análisis más clásicos. Mediante el algoritmo K-means y noveles técnicas de agrupamiento es posible obtener las regiones con mayor nivel semántico para adaptarlo al resumen de un documento. La eficacia del método se constata mediante el análisis de la entropía del resumen resultante.

[13] Teoría desarrollada en la teoría de autómatas

Los trabajos de Yuan y Sun (Yuan and Sun, 2004) se basan en las potencialidades de los algoritmos de agrupamiento para la categorización de grandes conjuntos textuales para descubrir en ellos nuevo conocimiento. Según los autores (Yuan and Sun, 2004) la mayoría de los métodos de agrupamiento se centran en la medición de la distancia o la similaridad, olvidando la estructura de los términos en los documentos. Esto hace que en esta investigación se presente un nuevo método denominado similitud coseno estructurado, capaz de suministrarle al documento una nueva forma de agrupar los términos a partir de su estructura y con el concurso de una ontología con el fin de mejorar la calidad del agrupamiento.

En Las investigaciones de Popescu (Popescu et al., 2004) se aborda el problema de la construcción de un resumen funcional a partir de la lógica borrosa, para obtener información sobre productos de información genética que se encuentran en una base de datos, cuyas anotaciones se realizan mediante una ontología y aplicando la lógica borrosa.

En los estudios de (Chen and Verma, 2006) abundan los trabajos a partir de UMLS (Unified Medical Language System), una ontología de la Biblioteca Médica Nacional de los Estados Unidos, donde se estudian a los usuarios que usan la ontología para servir de sustento a la construcción léxica. La investigación consiste en consultar la ontología y detectar determinados conceptos que sirvan de base a la summarización.

Las aportaciones de Yager y Petry (Yager and Petry, 2006) se centran en el entramado conceptual que facilite que los usuarios de los sistemas ontológicos sean capaces de controlar y manejar datos para descubrir en ellos conocimiento, aspectos básicos para enfocar la calidad de los resúmenes hacia las categorías diversas de una ontología. Los autores centran su investigación en las potencialidades del modelado de conjuntos difusos a partir de cuatro elementos implícitos y necesarios para un resumen: pertinencia, concisión y utilidad. Estos tres criterios permiten la construcción de una función de agregación que permite obtener una medida de calidad global capaz de sustentar la calidad de los datos

representados en el resumen. El desarrollo de la teoría se presentó mediante el análisis de un caso simple o de un solo atributo para delinear con claridad las cuestiones básicas y el enfoque, seguidamente este proceder se realizó con varios atributos, finalmente se logró obtener un resultado de datos más apegado a la necesidad pragmática del usuario (Yager and Petry, 2006).

Otros estudios se basan en el resumen de un léxico para la construcción de una ontología de dominio. En este trabajo, el léxico conceptual es representado por los sentidos que aparecen en la ontología de WordNet y sus relaciones. En la investigación cada frase de un documento está representada por un conjunto de sentidos de WordNet y constituye una transacción difusa para la minería léxico conceptual y la pertinencia de clasificación. Existe una versión de este sistema de generación automática de resúmenes , al igual que un método de evaluación intrínseca que facilite medir la calidad de resumen y la recuperación de la información (Hsun-Hui and Yau-Hwang, 2007).

En la extracción de video han aparecido métodos que explotan las ontologías como herramientas de anotación de video, debido a que estos han cobrado una elevada popularidad en los entornos Web sobre todo en los trabajos de Jin-Woo (Jin-Woo et al., 2007). La anotación automática de video mediante ontologías es presentada con el objeto de aprovechar las capacidades de la anotación en los procesos de extracción y recuperación de imágenes digitales, aplicando reglas de inferencia centradas en los conceptos siguientes: alto nivel de disparo / grupo / escenario / nivel de vídeo. Mediante una métrica se demuestra la calidad del método propuesto (Jin-Woo et al., 2007).

.Havents (Havens et al., 2008) presenta una ontología capaz de auto-organizarse para producir la visualización de la información que existe en un resumen. La auto-organización ontológica se desarrolla a partir de mapas que facilitan la visualización de información en los resúmenes especializados en genética (la ontología se muestra como una herramienta bioinformática y se denomina Gene Ontology (GO). La Organización de  (OSOM) consiste en una modificación de las herramientas de auto-organización desarrollada por Coñeen,

donde los datos permiten que la ontología sea generada a partir de las medidas de similitud de términos basadas en métricas de distancia.

Se presenta también en la sumarización de textos con ontologías un enfoque centrado en la extracción de pesos, los cuales son asignados a las clases y las subclases de una ontología jerárquica. Considerando los atributos de la ontología los autores (Henning et al., 2008) son capaces de mejorar la representación semántica del contenido de determinadas frases en el resumen.

Los autores (Henning et al., 2008) sostienen la teoría de que los clasificadores de frases asignan generalmente pesos a las frases en la taxonomía y son utilizando como motores de búsqueda, posición que no llega a ser totalmente fiable y difiere muchas veces de las necesidades del dominio.

Los autores (Hennig et al., 2008) utilizan un clasificador SVM capaz de identificar frases en el resumen mediante el uso de ontologías y la caracterización de de las oraciones. Los resultados experimentales muestran que la de extracción basada en ontologías supera la que se logra con los clasificadores de referencia, dando lugar a una puntuación más alta en la métrica Rouge cuando se analizan resúmenes extractos (Hennig et al., 2008).

Li y Motta (Li and Motta, 2010) han desarrollado un método ampliamente reconocido para facilitar la comprensión de una ontología y luego apoyar tareas diversas entre las que se encuentran: la reutilización de ontologías y la construcción de nuevas ontologías y resúmenes. En concreto, se investiga la aplicabilidad de las medidas de evaluación de las ontologías para la construcción de resúmenes acorde a diversas métricas y a la visión de los usuarios finales.

Otra postura sobre el trabajo ontológico aparece en (Ning and Shihan, 2006), quienes a partir de la ontología de tareas de la ONU desarrollan un conjunto de medidas de evaluación que facilite medir la calidad y el rendimiento de los sistemas de resumen, tanto a nivel de eficacia como de eficiencia, los resultados

de este proceso son evaluados acorde a las necesidades de los usuarios sobre el valor de los textos.

Finalmente en este análisis aparece el trabajo de Zhang (Zhang et al., 2007), donde se desarrolla un método para la selección de oraciones basado en una sentencia RDF como unidad básica para el resumen, para ello se propone caracterizar los vínculos entre las oraciones a partir de una ontología, en la cual se ejecuta una estrategia de reclasificación de oraciones. El estudio concluye con la verificación de la calidad de la ontología para la confección de resúmenes.

#### 2.10.1.2.- Mapas temáticos

Cartogramas sustentados en identidades, facetas y contextos que permiten la unión de entidades iguales aunque se denoten de forma diferente, facilitando el uso de consultas y filtros relacionados con un contexto o topic dentro de los sistemas.

**Topic maps**

El Origen de estos sistemas de visualización tiene como génesis el grupo de Davenport. Un espacio donde confluyen productores de libros electrónicos, surgido a mediado de la década de los noventa del siglo pasado. En el año 1993 se propuso la creación de una norma cuyo principal objetivo fuera posibilitar la fusión de índices impresos. Posteriormente evolucionó hacia otras estructuras (como tesauros), hasta llegar a ser una herramienta considerada en la Web para la organización, representación y gestión del conocimiento. La primera versión oficial del estándar ISO/IEC data del año 2000.

Es un conjunto de documentos expresados en un lenguaje que puede ser SGML o XML o una combinación de ellos interrelacionados en un espacio multidimensional donde las relaciones son precisamente las temáticas o topic. Conceptualmente la raíz del vocablo topic es la de temas y la representación eficiente de estos se da a partir de las relaciones o conceptos que tenga esa cartografía.

Los Topic Maps son herramientas para normalizar diversos elementos que

deben ser recuperados en la red. Para ello utilizan una notación especial que permite estructurar la información mediante el apoyo de una red de enlaces semánticos que relacionen diferentes recursos Informativos (ver figura 7). Los elementos principales de los Topic Maps son los topics (temas), las occurrences (la forma en que puede verse esa asociación) y las *associations*.

Esto en lo referente al resumen automático reviste una gran posibilidad de visualizar los recursos que existen en la bases de conocimiento o en los diccionarios, pues los elementos que la componen deben estar conectados y al efectuarse la búsqueda sobre estos se abre una posibilidad de establecer cartografías de alto nivel de especificación.

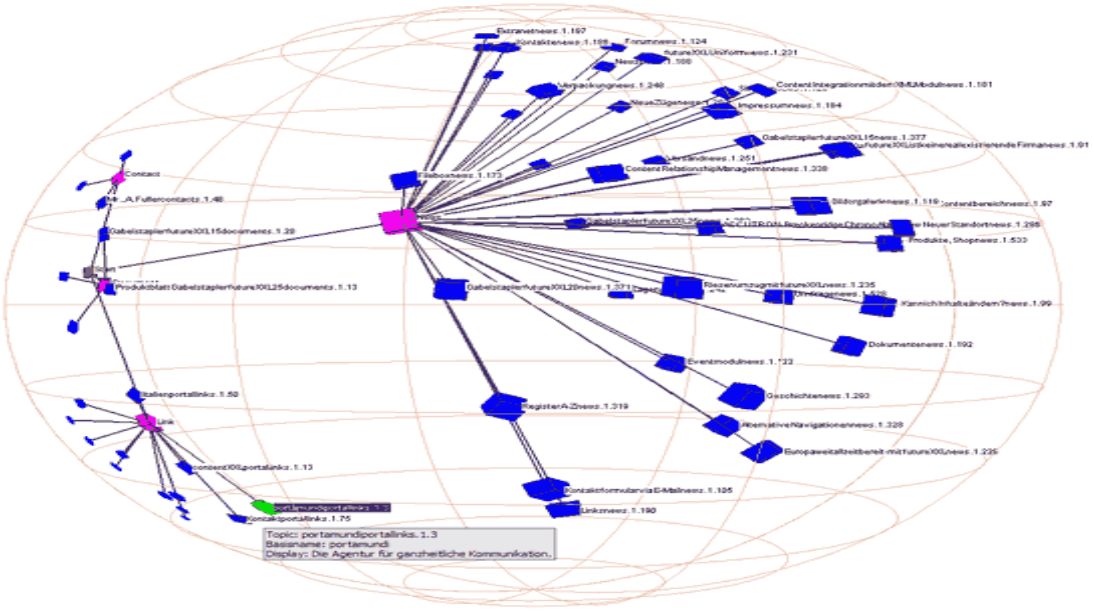



Figura 7 Topic Maps (Shaff, 2007)

### Mapas Conceptuales

Los mapas conceptuales son un medio de visualizar conceptos y relaciones jerárquicas entre conceptos. Se aprovecha la percepción como herramienta de comprensión pues se ha demostrado en trabajos de Kissil (Kissil, 2006) que el

recuerdo de imágenes visuales ayuda la comprensión de los fenómenos. Por tanto con la elaboración de mapas conceptuales se aprovecha la capacidad humana de reconocer pautas en las imágenes para facilitar el aprendizaje y el recuerdo.

Para Roque (Roque, 2008) los mapas conceptuales tienen por objeto representar relaciones significativas entre conceptos en forma de proposiciones. Una *proposición* consta de dos o más términos conceptuales unidos por palabras para formar una unidad semántica.

Un mapa conceptual es, por tanto, un recurso esquemático para representar un conjunto de significados conceptuales incluidos en una estructura de proposiciones que tiene por objeto representar las relaciones significativas entre los conceptos del contenido (externo) y del conocimiento del interno.

Constituyen un instrumental que facilita la representación del conocimiento de forma concisa ayudando a la transmisión de los mensajes conceptuales en toda su complejidad adoptando forma de grafos. Estos describen lo que se puede llamar mapas mentales, pues su acción sobre la percepción es una sinapsis del sistema neuronal. Los mapas mentales son herramientas que permiten la memorización, organización y representación de la información con el propósito de facilitar los procesos de aprendizaje, administración y planeación organizacional así como la toma de decisiones. Lo que hace diferente al Mapa Mental de otras técnicas de ordenamiento de información es que nos permite representar nuestras ideas utilizando de manera armónica las funciones cognitivas de los hemisferios cerebrales. Esta técnica organizativa fue descrita por Buzán (Buzan, 2003) con el objeto de describir diversas conexiones sinápticas que tienen lugar entre las neuronas de la corteza cerebral y que hacen posible prácticamente todas las actividades intelectuales del ser humano.

Su unión con el mapa conceptual es que representa estructuras de pensamiento, una herramienta indispensable para construir resúmenes y organizar cocimiento estructurado, lo que constituye incluso, una posibilidad

para desarrollar estrategias cognitivas que permitan establecer ergonomía en a la hora de crear un sistema de representación de la información (ver figura 8).

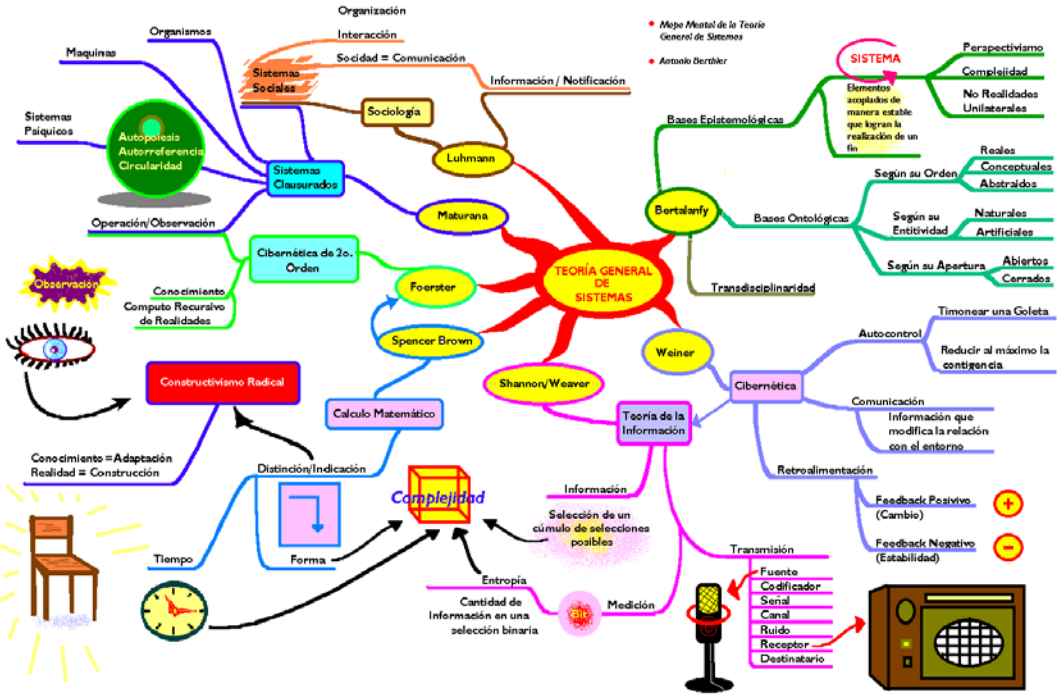



Figura 8. Mapa Mental (Buzán, 2003)

### 2.10.1.3.- Taxonomías

Según Hernández (Hernández, 2006) y Peralta (Peralta et al., 2006) el término procede del vocablo griego taxis lo que significa dirección. Estos recursos necesariamente nos llevan a las posiciones de las ciencias biológicas que utilizaron este vocablo para segmentar especies. En la Ciencia de la Información las taxonomías están vistas a través de la denotación sintáctica y relacional de los elementos de una disciplina o materia del conocimiento en forma jerárquica. Las taxonomías tienen una operatividad horizontal lo que las asemeja con los Sistemas clásicos de Clasificación que utilizan taxones o subdivisiones específicas, buscando conexiones lexicales, no sólo con el objetivo de organizar documentos representados en un espacio físico, examinando una posición

directa que facilite el interacción con el usuario. Las taxonomías son elementos condicionadores que facilitan la elaboración de conocimientos y la representación de contenidos, se conocen como estructuras relacionales o de posición. Su propia estructura facilita las relaciones jerárquicas y lingüísticas. Sin embargo, las taxonomías son elementos de los lenguajes documentales del siglo XIX. La propia estructura jerárquica del Dewey y la CDU tienen elementos taxonómicos (Peralta et al., 2006). Esto nos hace pensar que ya en el siglo XIX en la Documentación existían los taxones dados por la propia posición positivista de los preparadores del sistema, por tanto las taxonomías no son cualidades nuevas en los sistemas documentales, son recursos lingüísticos adaptados al presente de nuestra lingüística. Estos léxicos no operan por sí solos, lo hacen a través de posiciones que fueron luego las clasificaciones jerárquicas.

Las clasificaciones jerárquicas tienen una operatividad media en una taxonomía, habrá que buscar formas para elaborar fórmulas semánticas que enlacen todos los elementos que se disponen de forma vertical. Un modelo facético hecho de la forma en que Ranganthan concibió su sistema será necesario para el desarrollo y operación de las taxonomías. Estas no pueden limitarse sólo a la representación de los contenidos y a facilitar su búsqueda tienen que ser agentes de asociación de inferencias.

Necesariamente deben desarrollarse lenguajes de una poderosa posición asociativa, hecho al cual se encaminan; sin embargo, en cuestiones prácticas hoy poseen la limitación de sólo buscar una representación específica de una entidad determinada, una asociación basada en términos surgidos al calor del desarrollo de una disciplina, lo que permite ver sus asociaciones /estratos, pero no su proyección comunicativa, pues su campo de acción sólo está vinculado a públicos especializados.

Web Semántica

El desarrollo de Las Web como mecanismo de información y conocimiento ha traído a la humanidad avances importantes en lo referente a la producción,

utilización y evaluación de recursos informativos. Hoy se hace imposible combinar nuestras necesidades de búsqueda a través de esta amplia red (donde los contenidos están totalmente dispersos) y adaptar su uso para la búsqueda empleando nuestro idioma como interfaz ante un ordenador (Senso and Leiva, 2008). Por tal motivo surge la Web semántica.

Esta posición enfrenta sus criterios ante disímiles áreas de aplicación: mapas temáticos, mapas de facetas, XFML, RDF, etc. Es semántica esta estructura de Web ya que basa sus acciones a través de las estructuras de significado y la representación automática de las descripciones de metadatos. Hoy en día operan gracias a la inteligencia artificial sin olvidar su relación eminente con la lingüística, la semiótica y la Ciencia de la Información. Las Web o estructuras nodales de formatos World Wide Web poseen una organización de capas que permiten su operatividad. La Web Semántica no es más que una nueva representación de la Web actual, se trata de un espacio donde la información tiene un significado bien definido, de manera que pueda ser interpretada tanto por agentes humanos como por agentes computarizados.

Las estructuras que sirven de posicionamiento a la Web semántica son (facet maps, Topic maps, taxonomías) las cuales son capaces de brindar moldeabilidad a la Web ya que le proporcionan nuevas interacciones conceptuales, nuevas entidades que extienden la capacidad y los objetivos del sistema.

En la Web semántica se trabaja con Redes Semánticas o estructuras nodales basadas en la combinación de varias páginas Web, estructuras que tienen una fundamentación psicológica muy sólida, puesto que ponen de manifiesto el comportamiento que asume el usuario de caras a ese contenido con el que necesariamente interactúa. Los componentes básicos que encontramos en todas las redes son:

1. Estructuras de datos en *nodos*, que representan conceptos, unidas por *arcos* que representan las relaciones entre los conceptos.

2. Un conjunto de procedimientos de inferencia que operan sobre las estructuras de datos.

Las Web Semántica es un espacio donde la información tiene un significado bien definido, de manera que pueda ser interpretada tanto por agentes humanos como por agentes computarizados. En la Web estos sistemas (ontologías, facet maps, etc.) están diseñadas para ocupar el lugar de los tesauros automatizados definidos por estructuras rígidas y poco asociativas pero para que este rol pueda cumplirse se hace necesario un léxico más específico que aporte singularidad de posturas o sea, divergente de normativas como: ISO. Esto ha traído como consecuencia la aparición de lenguajes asociativos que permiten la relación de contenidos con mayor precisión que los estándares ISO sobre tesauros. Con ese fin se han creado los lenguajes diversos que permitan relacionar y/o asociar contenidos.

Las ontologías y todas las herramientas lingüísticas necesitan un lenguaje formal, un sistema de signos convencionales para ser representadas o construidas. La inteligencia artificial ha puesto a disposición de los trabajadores de la información diversas variables léxico-semánticas basadas en la lógica, que operan con predicados, realizando un análisis sintáctico como el KIF y C y facilitadores de modelos y de propuestas de casos. Existen otros que consideramos más asociativos, pero menos analíticos, por tanto los lenguajes de computación tendrán que tener una carga de inferencia, asociación y análisis importante.

### *2.11.- Retos de los sistemas de representación textual ante el resumen automático*

Las nuevas formas de representación documental tienen ante sí el reto mayor en los nuevos sistemas de recuperación y representación de información. Hay que tener en cuenta que la posición que ocupan hoy los documentos digitales en el mundo es preponderante. Nos enfrentamos a la necesidad de reponer y proponer estructuras léxicas que permitan la comunicación con el sistema; esta

posición no puede buscarse sólo desde el plano de actividades concernientes a los documentalistas, por tanto la asunción de posturas multidisciplinares y transdisciplinares es vital para desarrollar un sistema donde converjan estructuras de varias disciplinas. La humanidad se debate hoy ante la infoxicación, por tanto la posición que juegan estos sistemas de representación es importante.

Las propias estructuras organizativas derivadas del siglo XIX sirven de punto de partida a sistemas de estructura asociativa generados sobre una base cada vez más ergonómica buscando elementos de usabilidad, esto hace que se necesiten léxicos capaces de usarse de forma sencilla y que a la vez sean interactivos (Peralta, et.al. 2006). Para ello se demandan diálogos basados en estructuras semánticas con alto grado de inferencia que faciliten la cognición. Tiene que ser nuestra representación de información más ergonómica y predictiva, capaz de sintetizar las posiciones diversas del usuario, representativas de los diferentes casos de análisis sustentados en una dinámica que refleje la postura cognitiva sobre la que necesariamente interactúan los que usan la información.

Necesaria será también la flexibilidad, no se puede seguir construyendo sistemas de lenguajes sobre gramáticas complejas y multiestructurales (Hernández, 2006). Estamos llamados a construir sistemas flexibles capaces de reflejar actitudes y posiciones acorde al estado en que se encuentra el individuo. Un mundo virtual donde se inserte la lingüística y la representación necesita abordar lenguajes de consistencia que no se postulen como una ventana para enfrentar casos aislados, sino que enfrenten cognitivamente el universo comunicacional de las estructuras Web. Los léxicos generados y descritos en los procesos sumistas deben resolver el problema de la temporalidad, es decir asumirse desde una pragmática de comunicación que facilite la recuperación de la información. Se necesita utilizar las bases cognitivas generadas en los procesos léxicos, pues están sustentadas en el objeto real de las representaciones. Para ello se demandan modelizaciones de dominios claramente explicitadas con las unidades de representación y sintaxis

clarificadas al igual que sus formas gráficas y conectadas a procesos heurísticos de elevado nivel metacognitivo.

No se trata tampoco de eliminar las formas de representación clásicas del siglo XIX, sino de aprovechar su operatividad en nuestros sistemas actuales. Las herramientas que utilizan la Web semántica y el sistema de redes no pueden ser sólo estructuras lógicas (lenguajes de Programación), tienen que ser medios para una representación dinámica, que obligatoriamente tiene que estar sumida a normas y procedimientos.

### *2.12 El resumen automático: retos investigativos*

Como se aprecia en este análisis teórico, las investigaciones acerca del tratamiento de un resumen textual global[14] no alcanzan aún niveles de aplicación, al menos a nivel pragmático. La poca cooperación entre la Ciencia de la Computación y la Ciencia de la Información, así como el desconocimiento que tienen de este tema otras disciplinas ha retardado la presencia de estudios teóricos referentes a la solución de problemas referidos al tratamiento metodológico del resumen automático. Existen muy pocos estudios referidos a este apartado dentro de la Ciencia de la información, situación motivada por un lado, por las trabas que se dan para integrar las disciplinas necesarias para estudiar este campo y, por otro, por la apatía que sienten algunos profesionales de nuestra actividad cuando de representar información se trata. La Ciencia de la computación aún se debate entre la solución de los problemas de coherencia de los textos generados en los entornos de redes y desconoce otros elementos de la hipermedia, sólo se dedica al análisis de uno de sus elementos: el texto verbalizado. Los modelos automáticos desarrollan el proceso sumista intentando calcar los algoritmos mentales de los humanos y declarando estrategias cognitivas superficiales, es decir intentan ordenar la secuencia en que el hombre sintetiza y enuncia los textos.

[14] Resumen con imagen, sonido y texto, etc. En este apartado aparecen trabajos de Castell, Moreiro, Pinto, Vianello, Negroponte, Valle, Tramullas, etc.

Para resumir la hipermedia se necesitan otras lecturas del fenómeno, pues los métodos descritos (a pesar de desarrollar el trabajo con segmentos textuales que pueden estar o no dentro de la hipermedia) no llegan a desarrollar extractos donde se expongan de forma coherente todos los elementos que resumen la calidad de los textos.

Según Moreiro (Moreiro, 2004) los métodos que deban construirse para resumir necesitan tratar los aspectos siguientes:

Abordar nuevas visones teóricas que proporcionen noveles integraciones donde el análisis se convierta en un modelo puro no un híbrido con la mente humana

Transformar el resumen textual en resumen hipermedial, logrando que los extractos sean capaces de representar cabalmente todas las partes del texto. De esta forma se recomponen estructuras con cualidades semejantes a las estructuras textuales fuentes.

Desarrollar y mejorar la coherencia y el balance entre los textos generados en el hipertexto.

Describir mejores herramientas que permitan a los ordenadores la identificación de diversas cargas de información.

Generar resúmenes que sirvan de plantillas para la confección y unión de textos de forma coherente.

Avanzar en diversos aspectos como: tratamiento morfológico, lexicográfico, sintáctico (en los modelos algorítmicos), así como la aplicación de los conocimientos en redes semánticas.

Construir métodos donde los procesos inferenciales logren desbloquear lo que no se explicita como: el contexto, lo no dicho, las anáforas. Estos métodos también deben explicar la situación comunicativa y las competencias de los dominios.

Facilitar la construcción de textos hacia las características que enuncia la actividad documental hoy día. Es decir el logro de textos globales donde la

naturaleza del original no incida en su calidad, nos referimos con esto a la presencia en los textos de imagen, texto, sonido y voz,

Reconocer la importancia de los medios electrónicos como un espacio donde se mueven los documentos hipermediales.


# Referencias Bibliográficas

## 2.13. - Referencias Bibliográficas

AKAIKE, H. 1974. A new look at the statistical model identification. . *In:* IEEE (ed.).

AMAT, N. 1988. *Documentación científica y nuevas tecnologías de la Información,* Madrid, Pirámide.

AMIGÓ, E. 2005. QARLA: A framework for the evaluation of text summarization systems. *In Proceeding of 43rd Annual Meeting of the Association for Computational Linguistics.* Michigan.

ANSI 1978. *American National Standard for writing abastracts,* NuevaYork, ANSI.

ARCO, L. 2007. *Corpus miner: herramienta para el etiquetado de grupos y la obtención de extractos.* MSc., Universidad Central de las Villas.

ARCO, L. 2008. *Agrupamiento basado en intermediación diferencial.* Tesis Doctoral, Universidad Central"Marta Abreu" de las Villas.

BALDWIN, B. 2000. An Evaluation Road Map for Summarization Research. *In Proceeding of DARPA's TIDES (Translingual Information Detection, Extraction, and Summarization).* Washigton.

BALDWIN, B., DONAWAY, R., HOVY, E., LIDDY, E., MANI, I., MARCU, D., MCKEOWN, K.MITTAL, WHITE, M. V., MOENS, M., RADEV, D., SPARCK-JONES, K., SUNDHEIM, B., TEUFEL, S. & WEISCHEDEL, R. 2000. An Evaluation Road Map for Summarization Research. *The Summarization Roadmap*.

BANERJEE, A., KRUMPELMAN, C., BASU, S., MOONEY, R. & GHOSH, J. 2005. Model based overlapping clustering. *In Proceeding of International Conference on Knowledge Discovery and Data Mining (KDD).*

BARNERS-LEE, T., HENDLER, J. & LASSILA, O. 2001. The semantic web. *Scientific American magazine,* 284**,** 34-43.

BEZDEK, J. & PAL, N. 1995. Cluster validation with generalized Dunn's indices. *In:* KASABOV, N., COGHILL, G. (ed.) *In Proceeding of 2nd International two-stream Conference on ANNES.* Piscataway, NJ: IEEE Press.

BOCK, H. 1985. On significance tests in cluster analysis. *J. Classification***,** 77-108.

BORKO, H. & BERNIER, C. 1975. *Abstracting concepts and methods,* Nueva York, Academic Press.

BORKO, H. & CHATMAN, S. 1963. Criteria for aceptable abstracts: a survey of abstracters' instructions. *American Documentation*.

BRON, J. & DAY, M. 1993. *Quality in abstracting or summarization,* New York, MC-Graw –Hill.

BRUN, M., SIMA, C., HUA, J., LOWEY, J., CARROLL, B., SUH, E. & DOUGHERTY, E. R. 2007. Model-based evaluation of clustering validation measures. *Pattern Recognition,* 40**,** 807-824.

BUZAN, T. 2003. *Mental Maps in knowledge,* Boston, Atterville.

CALINSKI, R. & ARABAS, J. 1974. A dendrite method for cluster analysis. *Com.Statistics***,** 1-12.

CARLETTA, J. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics,* 22.

CORNELIUS, I. 1997. Information and Interpretation. *In Proceedings of COLIS 2.* Copenhague: The Royal School of Libraiansahip.

CREMMINS, E. 1985. *El arte de resumir,* Barcelona, Mitre.

CHACÓN, I. 2006. *La Mediación documental* [Online]. [Accessed 5.may. 2006].

CHAUMIER, J. 1986. *Análisis y lenguajes documentales,* Barcelona, Mitre.

CHEN, P. & VERMA, R. 2006. A Query-based Medical Information Summarization System Using Ontology Knowledge. *Proceedings of the*

*19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).* IEEE.

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DAUDEN, M. 1982. *Redacción de documentos,* La Habana, Pueblo y Educación.

DAVE, R. 1996. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters,* 17**,** 613-623.

DIXON, M. 1997. *An Overview of Document Mining Technology* [Online]. Available: Disponible en:http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dm.html [Accessed febrero.27. 2007].

DRÜSTELER, J. 2002. Information visualisation, what is it all about? *Inf@Vis! ,* 100.

DUNN, J. 1974. A fuzzy relative isodata process and its use in detecting compact well- separated clusters. *J. Cybernetics,* 3**,** 32-37.

FALKOWSKI, T., BARTELHEIMER, J. & SPILIOPOULOU, M. 2006. Community Dynamics Mining. *In Proceedings of 14th European Conference on Information Systems.*

FUKUMOTO, J. 2003. Text summarization based on itemized sentences and similar parts detection between documents. *In Proceedings of Third NTCIR Workshop.*

GOODMAN, L. & KRUSKAL, W. 1954. Measures of associations for cross-validations. *J. Am. Stat. Assoc.,* 48**,** 732-764.

HALKIDI, M., BATISTAKIS, Y. & VAZIRGIANNIS, M. 2001. Clustering algorithms and validity measures. *13th International Conference on Scientific and Statistical Database Management.* IEEE Computer Society.

HARMAN, D. & MARCU, D., ED. 2001. *Proceedings of the 1st Document Understanding Conference.* New Orleans, L.A.

HARTER, J. & BUSHA, L. 1990. *Metodología de la investigación en bibliotecología y Ciencia de la Información,* La Habana, Editorial Félix Varela.

HAVENS, T. C., KELLER, J., POPESCU, M. & BEZDEK, J. C. 2008. Ontological Self-Organizing Maps for Cluster Visualization and Functional Summarization of Gene Products using Gene Ontology Similarity Measures. *EEE International Conference on Fuzzy Systems (FUZZ 2008).*

HENNIG, L., UMBRATH, W. & WETZKER, R. 2008. An Ontology-based Approach to Text Summarization. *EEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* Princenton: IEEE.

HERNÁNDEZ, A. 2006. indización y Resumen. La Habana: Universidad de la Habana.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

HÖPPNER, F. 1999. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition West Sussex,* England, John Wiley & Sons

HU, P., HE, T., JI, D. & WANG, M. 2004. A Study of Chinese Text Summarization Using Adaptive Clustering of Paragraphs. *Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04.* ACM.

HUBERT, L. & SCHULTZ, J. 1976. Quadratic asignment as a general data-analysis strategy. *Br. J. Math. Stat. Psicol,* 29**,** 190-201.

HSUN-HUI, H. & YAU-HWANG, K. 2007. Towards auto-contruction of domain ontology : an auto-constructed domain Conceptual lexicon and its application to extractive summarization *Proceedings of the Sixth*

*International Conference on Machine Learning and Cybernetics.* Hong Kong: IEEE.

ISO 1976. *Documentation. analyse pour les publications et ladocumentation,* Ginebra, ISO.

JACKSON, P. & MOULINIER, I. (eds.) 2002. *Natural Language Processing for Online Applications* John Benjamins Publishing Company.

JAIN, A. & DUBES, R. 1988. *Algorithms for clustering data,* Englewood Cliffs, NJ, Prentice Hall College Div.

JAIN, A., MURTY, M. N. & FLYNN, P. J. 1999. Data clustering: a review. *ACM Computing Surveys,* 31**,** 264-276.

JIN-WOO, J., KYUNG-WOOK, P., JEONG HO, L., YOUNG SHIK, M., SUNG HAN, P. & DONG-HO, L. 2007. OLYVIA : Ontology-based Automatic Video Annotation and Summarization System using Semantic Inference Rules1 IEEE.

KARYPIS, G., HAN, E. & KUMAR, V. 1999. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling. *IEEE Computer,* 32**,** 68-75.

KAUFMAN, L. & ROUSSEEUW, P. 1990. *Finding groups in data: an introduction to cluster analysis*, John Wiley.

KIM, D. & PARK, Y. 2001. A novel validity index for determination of the optimal number of clusters. *IEEE Trans. Inform. Syst., E84- D***,** 281-285.

KIM, M. & RAMAKRISHNA, R. 2005. New indices for cluster validity assessment. *Pattern Recognition Letters,* 26**,** 2353-2363.

KISSIL, J. 2006. Mapas conceptuales. CDICT, Universidad Central de las Villas.

KUNCHEVA, L. & HADJITODOROV, S. 2004. Using diversity in cluster ensembles. *In Proceeding of IEEE SMC.* Netherlands.

LANCASTER, F. 1990. *Indexing and abstracting in theory and practice*, University of Illinois, Graduate School of Library and Information Science.

LANCASTER, F. 1993. *Indización y resúmenes: teoría y prática*, Briquet de Lemos/libros.

LANCASTER, F. 1996. *El Control del vocabulario en la recuperación de la información.,* Valencia, Universidad de Valencia.

LAROCCA, J. & SANTOS, A. 2000. A trainable algorithm for summarizing news stories. *In Proceeding of 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD).*

LARSEN, B. & AONE, C. 1999. Fast and effective text mining using linear-time document clustering. *In Proceeding of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

LASSO, J. 1969. *Manual de Documentación,* Barcelona, Labor.

LEZCANO, L. 2002. Modelos de tratamiento de textos y agrupamientos. Santa Clara, Cuba: Universidad Central"Marta Abreu" de las Villas.

LI, N. & MOTTA, E. 2010. Evaluations of User-Driven Ontology Summarization. *In:* CIMIANO, P. & PINTO, H. S. (eds.) *EEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* Springer-Verlag

LIN, C. 2004. Looking for a Few Good Metrics: Automatic Summarization Evaluation — How Many Samples Are Enough? *NTCIR Workshop 4,.* Tokyo, Japón.

LIN, C. & HOVY, E. 1998. Automatic Evaluation of summaries using n-gram co-ocurrence Statistic. *In Proceeding of HLTNAACL.* EE.UU.

LÓPEZ, J., COORD. 1996. *Manual de Información y Documentación,* Madrid, Pirámide.

LOWEIN, B. 2004. *Hipermedial discurse,* Canadá.

MANI, C., RAGHAVAN, P. & SCHÜTZE, H. 2008. *An Introduction to Information Retrieval,* Cambridge, Cambridge University Press.

MANI, I., GATES, B. & BLOERDORN, E. 1999. Improving Summaries by Revising Them. *In Proceeding of the 37th Annual Meeting of the ACL.*

MAÑA, M. 2003. *Generación automática de resúmenes de texto para el acceso a la información.* Tesis doctoral, Universidad de Vigo.

MARCU, D. 1999. Discourse trees are good indicators of importance in text. *In:* MANI, I., MAYBURY, M. (ed.) *Advances in Automatic Text Summarization.* MIT Press.

MATHIS, B., RUSH, J. & YOUNG, C. 1973. Improvement of automatic abstracts by the use of structural analysis. *JASIS,* 24**,** 101-109.

MAULIK, U. & BANDYOPADHYAY, S. 2002. Performance evaluation of some clustering algorithms and validity indices. *Pattern Anal Mach Intell,,* 24**,** 1650-1654.

MAYBURY, M. 1995. Generating Summaries from event Data. *Information Processing and Management* 31**,** 735-751.

MEDINA, J. & PÉREZ, A. 2007. ACONS: a new algorithm for clustering. *CIARP.*

MILLER, G. 2001. The W3C semantic web activity. *International Semantic Web Workshop.*

MILLIGAN, G. & COOPER, M. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* 50.

MOLONES, J. 2004. Introduction of semantic information in Information Science.

MOREIRO, J. 2004. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural,* Madrid, Ediciones Trea.

MORRIS, A., KASPER, G. & ADAMS, D. 1992. The efecets and limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research,* 3.

NAMBA, L. 1999. *Information Retrieval* Appleton.

NANBA, H. & OKUMURA, M. 2000. Producing More Readable Extracts by Revising Them. *Proceedings of the 18thInternational Conference on Computational Linguistics (COLING-2000).* Saarbrücken.

NEGROPONTE, N. 1995. *El mundo digital,* Barcelona, Ediciones B.

NELSON., T. 1992. *Literary Machines,* Sausalito, Mindful Press.

NING, H. & SHIHAN, D. 2006. Structure-Based Ontology Evaluation. *In: International Conference on e-business on Engineering.* Computer Society, IEEE.

NIU, Z., JI, D. & TAN, C. 2004. Document clustering based on cluster validation. *CIKM'04 Thirteenth ACM International Conference on Information and Knowledge Management.* ACM Press.

NORMALIZACIÓN, C. C. E. D. (ed.) 1983. *Resúmenes y Anotaciones (NC 39-12)* La Habana: CEN.

PAICE, C. & JONES, P. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. *16th International Conference on Research.* ACM Press.

PERALTA, M., LEIVA, A., ESTÉVEZ, V. & RUÍZ, M. 2006. *Retos y tendencias de la representación de la información y el conocimiento,* Santa Clara, Cuba, Editorial Samuel Feijoó.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez

POPESCU, M., KELLER, J. M., MITCHELL, T., BEZDEK, J. C. & 2004. Functional Summarization of Gene Product Clusters Using Gene Ontology Similarity Measures. IEEE.

PSICOLOGÍA, U. F. D. 2004. *Desarrollo y comunicación de documentos hipermediales.,* Santa Clara, Cuba, Editorial Samuel Feijoó.

RADEV, D., HOVY, E. & MCKEOWN, K., (EDS.) 2002. *Computational Linguistics (4)Special Issue on Summarization*, The MIT Press.

RITTBERGER, M. 1997. Measuring quality in the production of Databases. *Journal of Information Science,* 23.

ROQUE, M. 2008. Mapas conceptuales. Santa Clara, Cuba: UCLV.

ROSELL, M., KANN, V. & LITTON, J. 2004. Comparing comparisons: document clustering evaluation using two manual classifications. *IIn Proceeding of CON 2004 International Conference on Natural Language Processing* Hyderabad, India.

ROSMAN, C. 1996. *La Sociedad Digital y su perspectiva sociológica,* Santa Clara, Editorial Samuel Feijoo.

ROUSSINOV, D. & CHEN, H. 1999. Document clustering for electronic meetings: an experimental comparison of two techniques. *Decision Support Systems,* 27**,** 67-79.

SAGGION, H. & LAPALME, G. 2002. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics* 28**,** 497-526.

SALTON, G., ALLAN, J. 1997. Automatic Análisys Theme Generation, and Summarization of machina-readable Texts. *In:* SPARK, K., WILLET, P. (ed.) *Reading in Information Ridieval.* San Francisco: Morgan Kauffman.

SALTON, G., WONG, A. & YANG, C. S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,,* 18**,** 613-620.

SCHWARTZ, G. 1978. Estimation the dimension of a model. *Ann Statu,* 6**,** pp. 461-464.

SENSO, J. 2008. *Descripción e intercambio en la web semántica* [Online]. Granada: Universidad de Granada. Available: http://documentacion.ugr.es [Accessed 28. septiembre 2010].

SENSO, J. & EÍTO, R. 2004. Minería textual. *El profesional de la información,* 13**,** 11-27.

SENSO, J. & LEIVA, A. 2008. *Metamodelo para la extracción y desambiguación de textos científicos,* Santa Clara, Cuba, Universidad Central "Marta Abreu" de las Villas, Editorial Samuel Feijoó.

SETKIN, M. 2006. *Summarization,* Hanover, Atertile.

SHANNON, C. 1948. A mathematical theory of communications. *The Bell System Technical Journal of Artificial Intelligence Research,* 27**,** 379-423, 623-656.

SIEGEL, S. & CASTELLAN, N. 1988. *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill.

SILBERSCHATZ, A. & TUZHILIN, A. 1996. what makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Engineering,* 8.

SOLTEN, T. (2005). Sistema semiológico de la hipernmedia.

STEIN, G., BAGGA, A. & WISE, G. B. 2000. Multi-document summarization: Methodologies and evaluations. *TALN.*

STEINBACH, M., KARYPIS, G. & KUMAR, V. 2000. A comparison of document clustering techniques. *In Proceeding of KDD Workshop on Text Mining. Washington.* ACM Press.

TENOPIR, C. & JACSÓ, P. 2007. *Quality of abstracs* [Online]. Santa Clara. Available: Disponible en: http://intranet.cdict.uclv.edu.cu [Accessed 26.marzo 2007].

THEODORIDIS, S. & KOUTROUBAS, K. 1999. *Pattern Recognition,* Academic Press.

THOMPSON, S. & MANN, L. 1988. *Rethoric estructure in text*, McGraw-Hill.

TUZHILIN, A. 2002. *Handbook of Data Mining and Knowledge Discover,* Oxford, Oxford University Press.

VAN DIJK, T. 1978. *La Noticia como discurso: comprensión, estructura y producción de la información,* Barcelona, Paidós.

VAN DIJK, T. 1995. De la gramática del texto al análisis crítico del discurso. *BELIAR,* 2.

VAN SLYPE, G. 1990. *El servicio de documentación frente a la explosión de la información,* Buenos Aires, Consejo Nacional de Investigaciones Científicas y técnicas.

XIE, X. L. & BENI, G. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Learning,* 13**,** 841-846.

XU, G. 2004. *Data mining,* Londres, Oxford University Press.

YAGER, R. R. & PETRY, F. E. 2006. A Multicriteria Approach to Data Summarization Using Concept Ontologies. *IEEE TRANSACTIONS ON FUZZY SYSTEMS,* 14**,** 767-780.

YUAN, S.-T. & SUN, J. 2004. Ontology-Based Structured Cosine Similarity in Speech Document Summarization. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)* IEEE.

ZHANG, X., CHENG, G. & QU, Y. 2007. Ontology Summarization Based on RDF Sentence Graph. *WWW 2007.* Banff, Canadá: ACM.

ZANG, Z., HUANG, Z., ZHANG, X. & 2010. Knowledge Summarization for Scalable Semantic Data Processing. *Journal of Computational Information Systems,* 6**,** 3893-3902


# Capítulo III

## Los métodos de construcción automática de Resúmenes de Texto

# Capítulo 3. Los métodos de construcción automática de Resúmenes de Texto

## *3.1.- Introducción*

Los métodos de construcción de resúmenes fueron generados a partir de la década de los cincuenta del siglo *XX*. Son procedimientos de construcción sumista que incluyen modelizaciones estadísticas y semánticas. En estos procedimientos se tiene en cuenta las cualidades del texto y su relación contextual. La base constructiva de dichos procederes está anclada en diversos análisis entre los que pueden distinguirse los de dominio, semánticos y estadísticos. La finalidad de cada resumen es según Moreiro (Moreiro, 2004) identificar los elementos esenciales de un texto. Es incuestionable el valor que le ha conferido siempre la Ciencia de la Computación a los sistemas de minería de texto para el desarrollo de resúmenes.

Si bien, desde que aparecieron los primeros programas resumidores desarrollados por IBM se han dado saltos cualitativos en el tratamiento de los documentos textuales de forma automática, no es menos cierto, que los resultados en el campo de la lingüística computacional no son aún los esperados. La mayoría de los sistemas tiene como sustento algorítmico el análisis de las oraciones del texto, conseguido mediante el concurso de desambiguadores léxicos y mecanismos de tratamiento textual que le han dado a la Lingüística Computacional elementos necesarios para organizar y tratar los textos.

## *3.2.- Componentes de Un Sistema de Resumen*

Mediante el análisis de los sistemas de resumen de texto Moreiro (Moreiro, 2004) enuncia las siguientes consideraciones para la automatización de un texto:

**Entrada**:

- Superestructura del documento.

- Especificidad.
- Nivel de Especialización.
- Tamaño del Documento.
- Información Textual o Multimedia.
- Género. (Artículos Científicos, periodísticos, etc.)
- Unidad. (monodocumento o multidocumento)
- Lengua.

**Finalidad**

- Destino (Producto final o incluido en un sistema mayor)
- Destinatarios (adaptado o no a las necesidades de usuarios)
- Uso.

**Salida**

- Contenido (datos concretos o información remática)
- Formato.
- Tipo (informativo o con información agregada)
- Procesos de Producción (extracción o abstracción)
- Tamaño.

### *3.3.- La organización del Texto en los contextos digitales*

La estructura de los textos que coexisten en INTERNET tiene gran importancia para el desarrollo de las técnicas de tratamiento textual. Los distintos segmentos del texto facilitan la organización de los significados y mejoran el proceso de análisis automático. Las construcciones textuales poseen elementos que se combinan para formar categorías textuales superiores, es decir generan diversos niveles de texto los cuales permiten la operatividad de los sistemas informáticos. Los agentes usan la macroestructura del texto para organizar y jerarquizar las

oraciones (Hernández, 2007). Según Moreiro (Moreiro, 2004) la macroestructura constituye todo la información contenida en un discurso o parte de él, por tanto el análisis de la macroestructura facilita el trabajo con los métodos de corte semántico/cognitivo o sea, coadyuva al desarrollo de un conjunto de procedimientos denominados: Métodos de Estructura Profunda.

Van Dijk y Kintsch (Van Dijk and Kintsch, 1983) opina que el nivel de significación más profundo es la macroestructura y que sólo en ella se dan las cargas de significación más amplias. Es este segmento textual el que permite al agente reconocer el vínculo de todas las afirmaciones o frases que constituyen el texto. Los robots gracias a las estructuras textuales pueden leer las partes del texto de forma jerárquica hasta obtener la sección más general, lo que propicia que la información de un conjunto textual pueda ser expresada de forma coherente como unidad semántica. Para Amaro (Amaro, 1977) sin la macroestructura (ver figura 8) la coherencia de un texto sería superficial y lineal. Las mencionadas estructuras favorecen el análisis de las unidades semánticas del discurso, lo cual constituye uno de los fundamentos lingüísticos esenciales para el desarrollo de los sistemas de Procesamiento del Lenguaje Natural. Los sistemas de extracción de textos presentan diversos componentes que operan mediante el reconocimiento de las distribuciones del texto. Pinto (Pinto, 2001) enuncia la estructura de un sistema de resúmenes con los niveles siguientes:

- **Nivel de texto:** permite desarrollar o analizar los textos y descomponerlos.
- **Nivel de palabra:** analizadores sintácticos y morfológicos.
- **Nivel de frase:** desarrolla sus acciones a nivel de sintagma y diversas categorías de significados.
- **Nivel interoracional**: basa sus acciones en co-referencias.
- **Nivel de esquema**: proyección de los formatos de salida que permiten desarrollar diversas ofertas de información.

| | | | |
|---|---|---|---|
| **Estructura superficial** | **Microestructuras:** secuencia de oraciones concretas | | |
| **Estructuras intermedias** | **Macroestructuras:** estructura semántica parcial | | |
| | **Superestructura:** disposición de las partes del discurso | | |
| **Estructura global** | **Macroestructura global:** estructura semántica global, de dimensión | *Sintáctica* | • estructura temática transformacional,<br>• sintaxis de base lineal. |
| | | *Semántica* | • contenido global del discurso, su tema. |
| | | *Pragmática* | • función comunicativa;<br>• generación y comprensión textual (macrorreglas y estrategias);<br>• coherencia y cohesión del discurso. |

Figura 9. Unidades Semánticas del Discurso (Pinto, 2001)

### *3.4. Análisis de Dominio y Extracción de Texto*

Hay muchas investigaciones acerca del accionar de los usuarios, entre ellas se encuentran: el comportamiento en la búsqueda de información, los estudios de necesidades, las investigaciones de usabilidad, etc. Estas investigaciones se reportan en la literatura de la BCI (Bibliotecología y Ciencia de la Información), desde la época de Bernal (1948) citado por Núñez (Núñez, 2005). En los marcos ciberespaciales estos estudios toman preponderancia, pues a partir de ellos se desarrollan técnicas de modelación que permiten describir y construir modelos de usuarios para fines específicos.

Los enfoques sociocognitivos representan la clave fundamental de los sistemas de extracción de texto. En Hjorland (Hjørland, 2004) pueden apreciarse las dimensiones de los estudios de usuarios y su vinculación con eminentes figuras del mundo de la Ciencia de la Información como Garvey y Griffith (Garvey and Griffit, 2008) quienes desarrollan los primeros estudios de las llamadas comunidades epistémicas. La Ciencia de la Computación ha abordado los estudios de usuarios desde una perspectiva que evade algunas de las

cualidades de su comportamiento. Los estudios de dominio no pueden seguir siendo una suma de cualidades comunicativas. Los mecanismos utilizados para la construcción de software nos demuestran que existe un tratamiento diferente dentro de la Ciencia de la Computación para el análisis de dominio, lo que se da por la poca interdisciplinariedad que existe en los estudios sustentados desde la cibernética y el hermetismo en que vive la Ciencia de la Información.

Para Hjorland (Hjørland, 2004) la Ciencia de la Información, las Ciencias de la Computación y las Ciencias Cognitivas han descuidado los factores culturales, mediadores en las relaciones entre personas e información y sólo han intentado estudiar la relación generalizada entre las personas y algo denominado "información". Arco (Arco, 2007) opina que los agentes se presentan generalmente ante los usuarios para pretender ser reconocidos como herramientas de valor. Esta aseveración es válida, si se tiene en cuenta que en los sistemas que se describen no existe un total análisis de dominio. Se habla de análisis de discurso exclusivamente cuando se desarrollan técnicas que trabajan con la estructura textual, sin embargo, en los métodos que operan con la estructura superficial del texto – los más comercializables- no se evidencia el tratamiento de los dominios a los que se dirigen.

La tendencia de intentar medir la necesidad del usuario formulando preguntas o estudiando su comportamiento, no es suficiente para organizar conocimiento. La información que se necesita para solucionar un problema de organización del conocimiento donde necesariamente están implicados los humanos, no es una cuestión psicológica, sino teórica/filosófica (Hjørland, 2004). Para generar un nuevo modelo de resumen se necesita interpretar el comportamiento comunicacional de los usuarios. Los estudios de (Núñez, 2005) describen una triada indisoluble que busca reflejar la relación esencial entre metateoría, métodos y metodología que, según Hjorland (Hjørland, 2004), le da un peso importante a las interpretaciones de los estudios de usuarios.

En Andersen (Andersen, 2006) se han visto tratados que demuestran aplicabilidad en los estudios de usuarios en entornos digitales y en la selección

de recursos de información. El análisis de dominio es una herramienta válida en los entornos electrónicos, siempre y cuando los estudios se realicen a partir de métodos coherentes. Mediante esta técnica se pueden reconocer los elementos esenciales que facilitan los estudios de usuarios de una comunidad determinada. Esta dimensión investigativa expresada de formas bidireccionales abrirá las puertas para la implementación conjunta de los procesos esenciales en la producción de resúmenes.

### *3.5. Procesos fundamentales en la producción de Resúmenes*

La confección de resúmenes automáticos tiene dos vertientes fundamentales: extracción y abstracción.

La extracción trabaja a partir del análisis sintáctico de las oraciones y las frases. Esta forma de modelar sistemas es muy eficaz en la extracción de textos que no poseen una retórica específica, lo cual hace que esta variante de tratamiento textual sea muy apropiada para sistemas que tienen diversos dominios como clientes. El volumen de extracción con este método puede ser tal, que impida la calidad de los resúmenes al desarrollar problemas diversos en el orden de la cohesión[15], la coherencia[16] y el desbalance[17].

La abstracción opera a partir de análisis semánticos y es apropiada para usuarios específicos, pues con la referida técnica se obtienen mejores dividendos que la anterior en lo referente a la cohesión, el balance y la coherencia, al ser un método netamente semántico, propicia que se pueda trabajar con corpus detallados y formatos textuales de estructura retórica.

La existencia de una estructura o aplantillamiento en la forma de producir la información en diversos grupos de usuarios es de vital importancia para la construcción de resúmenes basados en abstracción. Diversas plantillas son descritas en la literatura por Gaizauslas-Wilks (Gaizauslas – Wilks, 1988) quien explica que los sistemas de extracción basados en plantillas son muy utilizados

[15] Es el orden lógico de los elementos textuales. Es presentar de forma lógica el texto
[16] Se refiere a la unidad del texto, es decir, es la conexión entre los elementos textuales.
[17] Una dificultad que se da por la falta de nivelación entre los párrafos del texto.

como métodos extractivos en los medios periodísticos, fundamentalmente en la prensa económica. Existen ejemplos de agencias como Reuter, que utilizan estos métodos y otras que realizan artículos de información financiera mediante la extracción automática de referencias bibliográficas de patentes a texto completo. Este tipo de procesamiento del lenguaje natural se cumplimenta con plantillas estructuradas en patrones.

El buscador Altavista es ejemplo del uso de plantillas en las que pueden verse frases que se relacionan con el tema permitiéndole al usuario la búsqueda de información por la opción que este decida. Según Radev (Radev, 2001) las plantillas Web tienen calidad y se basan sólo en el análisis y la síntesis oracionales, permitiendo la jerarquía de conceptos y su generalización.

Aunque abundan diversas técnicas de extracción automática que operan con resúmenes estructurados, la producción científica apunta hacia la difusión de procedimientos sustentados en la recuperación de información (activados por extracción), lo que se patentiza en diversos trabajos de autores como Sanderson (Sanderson, 2005), Spink (Spink, 1997) e Ingwersen (Ingwersen, 1992).

La visión de estos autores responde al planteamiento de la necesidad de modelar nuevas formas de estudio de necesidades enfocadas desde la Ciencia de la Computación, la cual asume posiciones teóricas que por sí solas no rebasan los marcos formales de la comunicación y conducen a un falso análisis del usuario, poniendo delante de la representación textual un modelo de recuperación informativa centrado en una disciplina denominada ingeniería de software, la cual es poco operativa para conocer los diversos mecanismos comunicativos de los futuros usuarios. Otros autores como Pinto (Pinto, 2001) y Mizzarro y Tasco (Mizarro and Taso, 2004) sostienen sus análisis de los métodos de extracción teniendo como centro del estudio los diversos discursos de los usuarios.

El análisis del discurso facilita el conocimiento de los emisores. Metodológicamente admite la organización de los segmentos discursivos, pues

facilita diversos criterios para el agrupamiento de unidades que representan los conocimientos específicos de determinado conjunto de personas. Van Dijk (Van Dijk, 1980) sostiene que su relevancia en el campo de la investigación se da por la capacidad que tiene de describir los dominios, pues como instrumento de investigación es una herramienta adecuada para producir descripciones explícitas y sistemáticas, tanto textuales como contextuales de las unidades del uso del lenguaje al que se denomina discurso.

Las dimensiones textuales se refieren a las estructuras del discurso en los diferentes niveles de descripción, mientras que las contextuales relacionan a estas con las propiedades del contexto. Esta herramienta de estudio de la lengua es competente para lograr la asimilación de mensaje producto de las formas de comunicación humana. Van Dijk (1980), citado por Zaldua (Zaldua, 2006) sostiene que los miembros de determinada comunidad poseen un lenguaje común, piensan e interactúan entre ellos, lo que les permite representar su cultura y su contexto por medio de la lengua, que es un reflejo de los estratos sociales en lo que se desenvuelve determinada comunidad epistémica.

Los planteamientos del análisis de discurso se han convertido en una especie de paradigma de análisis de la representación de dominios textuales, sin embargo, no se ha declarado qué elementos del análisis de discurso serán aplicados en los estudios de comunidades epistémicas con vistas a la confección de sistemas de resumen automático.

### *3.6.- Métodos de extracción de Texto Verbalizado*

Los métodos de construcción de texto se presentan aunados bajo diversos criterios. Matcalf (Matlclaf, 2006) se apela a la siguiente clasificación de los métodos agrupando las técnicas de construcción por métodos automáticos en tres grandes grupos:

- Técnicas basadas en la superficie del texto. (no se realiza ningún tipo de análisis lingüístico)

- Técnicas basadas en las entidades nombradas en el texto. (se realiza algún tipo de reconocimiento y clasificación del léxico utilizado)
- Técnicas basadas en la estructura discursiva. (requieren algún tipo de tratamiento estructural del documento, generalmente de tipo lingüístico).

D´cunha (D'Cunha, 2006) propone una clasificación donde los métodos se agrupan de la siguiente forma:

- Métodos superficiales.
- Métodos de nivel medio.
- Métodos profundos.

Por su parte María Pinto (Pinto, 2001) clasifica los métodos de construcción automática de resúmenes de la forma siguiente:

- Métodos de extracción basados en la estructura superficial.
- Métodos de sumarización. (a medio camino entre resumen y extracto)
- Métodos gráficos y relacionales.
- Métodos de extracción basados en la estructura profunda.

El autor selecciona la clasificación de María Pinto (2001) por estar más cercana al campo de la Ciencia de la información.

### 3.6.1. Métodos de extracción basados en la estructura superficial

Los métodos superficiales incluyen toman como referentes las frecuencias de palabras. En estos métodos se pone de manifiesto que las entidades más importantes tienden a ser mencionadas más a menudo, lo cual propicia que se le atribuya más peso a las oraciones que poseen un mayor número de palabras frecuentes inusuales. El pionero en modelar un método estadístico para resumir textos fue Luhn (Lunh, 1958), el que estableció una ponderación en las palabras que contenían los textos. Le otorgó valor cero a los términos vacíos y a los no vacíos que representaban un valor superior a una medida antes definida eran consideradas representativas del texto, rechazando las restantes. También creó

la posibilidad de medir la representatividad de las locuciones combinando la reiteración de los términos en el texto y la vecindad entre ellas.

Edmundson (Edmundson, 1969) tomó como núcleo la teoría de Luhn para desarrollar sus métodos. La originalidad de sus postulados fue según Moreiro (Moreiro, 2006) incorporar a las oraciones un nuevo factor de decisión. Edmundson (1969) estableció pautas para la orientación de las oraciones en el texto. Este proceder permitió calcular el peso de las palabras de acuerdo con la parte del documento donde apareciera (introducción, resultados, conclusiones) algo muy similar al resumen aplantillado. También esta metodología proponía el desarrollo y análisis de los textos a partir de las oraciones iniciales de los párrafos. Luhn (1958) y Edmundson (1969) consideraban la aparición de las palabras en los títulos, subtítulos y epígrafes de los documentos, ya que pensaban que estos elementos tenían mayor valor para la representación del contenido documental.

Dentro de esta clasificación de métodos aparece la subdivisión descrita por Paice (Paice, 1988)donde se enumeran los métodos estadísticos siguientes:

- Frecuencia de palabras: La selección de palabras clave realizada en atención a la frecuencia de repetición en la frase.
- Palabras clave del título: Selección sobre el título y encabezamientos, tras eliminar las palabras vacías.

Algunas variantes del método antes referido residen en la localización de la información por oposición en el texto, del mismo se encuentran las combinaciones siguientes:

- **Lead method**: Este procedimiento considera que en cualquier texto lo importante aparece al principio o al final, por lo cual su criterio de selección para los extractos son las primeras oraciones o párrafos.
- **Optimun position policy (OPP)**: Se acciona a partir de las estructuras que trabajan con cargas de significados que aparecen en diversas posiciones del texto dependientes del género. Lin (Lin and Hovy, 1998)

precisa que la eficacia de este método es mayor cuando trabaja a nivel oracional y de párrafo.

- **Sintáctico**: La extracción relacionada con la estructura sintáctica del texto y los elementos de las oraciones en los párrafos son utilizados como puntos de lectura y acceso. Esto permite su aplicación a dominios amplios. Aunque la calidad de los extractos no es la deseada se logran representar grandes volúmenes de información textual.
- **Palabras temáticas importantes:** Luhn, (1958), (Edmunson, (1969) y Kupiec (Kupiec, 2003) describen esta técnica de análisis estadístico de la estructura del texto que opera mediante la identificación de términos clave dentro de los documentos.
- **Términos o frases significativas** (cue phrases). La selección se apoya en una lista de términos indicativos que tienen un peso en la selección de las frases. Se diferencian tres clases según la relevancia: *Bonus phrases* (indicadores de términos con peso), *Stigma phrases* (indicadores de términos de escaso peso) y Null phrases (indicadoras de frases irrelevantes para la selección).

Los extractos textuales relevantes se logran mediante representaciones semánticas del texto. Edmundson (Edmundson, 2006) propone un método lógico-matemático para la asignación de valores numéricos a las frases. Sus basamentos muestran cuatro modos de selección de frases:

- palabras clave (key).
- entrada (cue).
- título.
- emplazamiento.

La evaluación de los experimentos permitió reconocer que la selección por entrada, título y emplazamiento tuvieron ventaja frente a los métodos de

valoración que se circunscriben sólo a palabras clave, esto propicia que este método sea ignorado en la construcción de resúmenes automáticos.

ANES (Sistema de Extracción Automática de Noticias) propone una combinación de métodos sobre la base de técnicas estadísticas/heurísticas sobre los términos, determinando las frases más representativas. El análisis estadístico del corpus documental se realiza a partir de las frecuencias, asignándoles un peso a cada término y un identificador de léxico. Gracias al uso de sistemas de ecuaciones, donde se analizan las frecuencias con que se seleccionan las frases que contienen las ideas principales representadas por los términos, unido a la suma de los pesos de los vocablos que contiene cada frase y enunciando los conjuntos de palabras que poseen los mayores pesos logra sus operaciones este sistema.

ANES logra grandes volúmenes de efectividad. Los resultados de esta técnica fueron analizados en diversas fases experimentales y fueron valorados por analistas con la calificación de medianamente aceptables en cuanto a coherencia y cohesión.

El método de Knowledge Discovery [18] (KD) es otra técnica descrita bajo esta clasificación, es un proceso previo por el que pasan los documentos electrónicos, la misma se lleva a cabo mediante la supresión de términos y datos el texto original, reduciendo el documento primario a una secuencia de palabras. Esta herramienta analítica reconoce el máximo de secuencias de palabras que propician la ordenación de los textos, lo que permite determinar una estructura para la clasificación y ordenación de términos.

Este método permite (dentro de un grupo de documentos) encontrar los términos más frecuentes en una colección y realizar una comparación contra colecciones de documentos mayores. KD es un método aplicado a técnicas de análisis inteligentes muy utilizados en bases de datos automatizadas, lo cual propone

[18] Método descrito en diversas publicaciones de Inteligencia Artificial

generar conocimiento de bases de datos reales, aplicando diversos patrones para la extracción de información.

El fenómeno de Internet trae aparejado para los usuarios diversos fenómenos dados por las características de la organización de información en estos entornos.

Como resumen de este apartado es importante destacar que estas técnicas fueron las más utilizadas cuando comenzaron a hacerse los primeros resúmenes automáticos y aún son muy utilizadas en la mayoría de los sistemas comerciales. El procedimiento elaborado por (Lunh, 1958, Edmundson, 1969) es la técnica más utilizada en los inicios de la sumarización, el mismo se basa en el supuesto de que la frecuencias de las palabras son un elemento importante para la selección de oraciones. Otra técnica desarrollada a partir de las ideas de Lunh (Lunh, 1958) es la denominada “Posición de fragmentos del texto”, generalmente utilizada para extraer texto con una retórica específica pues permite reconocer (títulos, apartados,...). Por ejemplo si se piensa en un artículo con determinada estructura retórica, las primeras líneas serán las más relevantes (Brandow et al. 1994, Lin y Hovy 1997, 1999)

Ha sido muy difundida la técnica que se denomina: Identificación de Palabras o Frases Clave (cue frases). Esta técnica se sustenta de la relevancia que indican algunas palabras o frases como por ejemplo, “se observa”… “se destaca” o “en conclusión”(Edmundson, 1969). Existen otras técnicas que analizan las palabras y las frases irrelevantes que han de ser eliminadas de cara al resumen como la de Tufel y Moens (Teufel and Moens, 2002).

Se han reportado en la literatura de consulta sistemas de resumen que se apropian de técnicas estadísticas para extraer la información que facilita la obtención de datos cualitativos. En Berger y Mital (Berger and Mittal, 2000) se describe un sistema capaz de resumir textos de páginas Web utilizando modelos probabilísticas basados en aprendizaje automático (Brnadow and Mitze, 2004), el mismo condensa noticias de forma automática para 41 publicaciones

especializadas en Comercio. Por su parte Dunning (Dunning, 1993) desarrolla un test basado en Log-likelihood Ratio Test , aplicando una distribución binomial en vez de una distribución normal.

Es importante en esta técnica también referir los sistemas que se basan en aprendizaje automático, entre los que se encuentran los siguientes:

- Kupiec (Kupiec et al., 1995) desarrolla un sistema de resumen automático que toma como sustento el entrenamiento, con un elevado basamento en la estadístico. En este mismo terreno aparece el método Optimal Position Policy, el mismo aplica una técnica capaz de aprender cuales son las oraciones con más peso en un corpus de entrenamiento (Lunh, 1958, Lin and Hovy, 1998).
- En Fuentes et al. (2003) aparece un sistema extractivo capaz de extraer la frase más relevante de un texto, utilizando una aproximación de aprendizaje automático, la cual se contrasta contra reglas manuales para obtener mayor calidad en el proceso.

Para finalizar el análisis de estas técnicas el autor considera que es vital declarar algunos sistemas puramente estadísticos que han dado resultados regulares, la mayoría están basados en el Vector Space Model (VSM) (Salton and McGillm, 1983), modelo que sustenta los sistemas CORTEX (Torres-Moreno et al., 2002) y el sistema ENERTEX (Fernández et al., 2007). CORTEX es un sistema de resumen capaz de construir textos a partir de algoritmos de decisión capaces de combinar varias métricas obtenidas del procesamiento estadístico y del VSM. ENERTEX está destinado a resumir información en el campo de la Física estadística, sus acciones se basan en la codificación del texto a partir de espines.

Estas dificultades pueden ser aminoradas mediante el empleo de herramientas de *sumarización.*

### 3.6.2.- Métodos de Sumarización

Los sistemas que cargan en su estructura lógica la sumarización están muy ligados a la estructura del discurso (retórica), pues incorporan cierto grado de creatividad y operan sobre elementos de construcción textual como: cohesión y coherencia. Dentro de la sumarización se distinguen dos grandes agrupaciones:

- Métodos basados en la estructura discursiva o retórica.
- Métodos basados en la estructura profunda.

Barzilay (Barzilay, 1988) sostiene que los sistemas basados en la cohesión, relacionan los ítems de un texto, la conectividad no estructural, la repetición y la referencia. En ellos la cohesión léxica se logra mediante la selección de las palabras. Según Marcu (Marcu, 1997) estas metodologías de tratamiento textual consiguen interrelacionar los segmentos del texto, pues operan con los elementos del discurso (trabajan a nivel de macroreglas y microreglas) y usan la estructura semántica a modo de conexión. Esta proposición metodológica basada en la coherencia interna del texto, consigue una buena calidad del resumen. Marcu (1997) representa la estructura retórica del texto y utiliza algoritmos de análisis retóricos basados en un corpus delimitado por marcadores de discurso y compuesto por fragmentos textuales.

La estructura profunda se erige como un método muy avanzado donde el tratamiento textual es el eje fundamental del proceso en el que se centra el agente. Estos y otros ejemplos reflejan una perspectiva desde donde abordar el reto del resumen automático, sin embargo, existen 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, 1997).

El autor enfatiza en que este aspecto será uno sobre los que se centrará su propuesta metodológica, ya que, como se verá, la representación discursiva de

los textos es una de las informaciones (entre otras, como la sintáctico-comunicativa) que debe integrarse en nuestro modelo de sistema de resumen automático. Para formular relaciones discursivas mediante esta técnica se necesitan marcaciones de discurso.

En el campo del procesamiento del lenguaje natural mediante un corpus de referencia se pueden formular diversos niveles de marcación como los siguientes:

- **Corpus codificados o etiquetados.** Se entiende por codificación de textos todos aquellos procedimientos de marcación destinados a obtener un texto caracterizado y preparado para la fase de análisis.
- **Corpus lematizados.** En los corpus lematizados cada palabra contiene una indicación de su lema (por ejemplo, en los verbos: el infinitivo; en los sustantivos y adjetivos: las formas de masculino singular). Un lema es una forma canónica elegida arbitrariamente, bajo la cual se agrupan el resto de las formas de su paradigma morfológico.
- **Corpus analizados**. Además de las etiquetas, un corpus analizado contiene una información de más alto nivel, como corchetes de marcación de constituyentes, estructuras arbóreas etiquetadas. El análisis puede realizarse en diferentes niveles de descripción lingüística: sintaxis superficial, sintaxis de estructura profunda, estructuras pragmáticas o discursivas.

Los métodos basados en la estructura profunda operan con un sistema experto basado en una red de conocimientos básicos o redes semánticas representativas del contenido y aplicados al texto. Se postulan como un modelo semántico conceptual de generación de sumarios basado en las teorías cognitivas. Todos los sistemas generados a partir de la estructura superficial del texto producen problemas de coherencia, aunque el desbalance textual es menor si se comparan con los métodos basados en la estructura superficial.

Es erróneo anclar el desarrollo de un extracto sólo en evidencias estadísticas, ellas no atestiguan la obtención total de las oraciones del texto. Es por ello que es necesario buscar métodos donde la semántica (lingüística) y las estructuras textuales estén identificadas.

### 3.6.3.- Métodos de sumarización basados en la estructura retórica

La estructura retórica está declarada a partir de sistemas de generación de resúmenes en los cuales se extracta información tomando como base la organización del texto original o fuente, ya que en esta forma textual pueden observarse secciones que comienzan con vocablos determinados, lo que facilita la detección de secciones en el acto del resumen. Pinto (2001) opina que los resúmenes obtenidos mediante este método tienen un alto nivel de coherencia, puesto que respetan la semántica del texto original.

Sobre este tipo de resumen se observan diversos trabajos entre los que se encuentran los de Marcu (1997) quien demuestra en sus estudios los nexos entre los árboles de estructura retórica y las unidades textuales reconocidas en los procesos de evaluación. El estudio analítico del árbol de estructura retórica se realiza mediante el desarrollo de marcadores discursivos o marcadores de discurso, también este estudio textual suele hacerse trabajando con frames o fragmentos textuales. Los analizadores retóricos identifican los marcadores discursivos del texto y las unidades elementales que lo definen en el esquema retórico y propician la detección de las unidades esenciales del texto.

#### 3.6.3.1.- Plantillas

En Gaizauslas-Wilks (Gaizauslas – Wilks, 1988) aparece una descripción del uso de las plantillas para hacer resúmenes y recuperar información. Está técnica desarrollada y descrita por los referidos autores, sólo es aplicada a entornos digitales, siempre y cuando la información se encuentre estructurada. El uso de las plantillas se da en áreas disímiles de la actividad humana en las que su utilización como medio extractivo no requiere del uso de fuentes textuales con características retóricas disímiles, lo que ayuda a la obtención de excelentes

resultados. Una de las áreas en la que más se desarrolla este tipo de método es en la extracción de noticias mediante el concurso de sistema como SCISOR [1990], JASPER [Reuter, 2003], los cuales emplean como input de datos las plantillas y las técnicas de análisis parcial. También existen otros sistemas como FIES orientado a la extracción de información financiera de artículos de prensa en entornos digitales. Otros campos como la Química y la extracción de referencias de patentes han trabajado con sistemas resumidores basados en plantillas.

#### 3.6.3.2.- Plantillas Web

La Web al ser uno de los medios esenciales de desarrollo y difusión de conocimientos en el espacio digital se ve muy ligada al desarrollo de la actividad del resumen ya que se erige como una herramienta esencial para el procesamiento y recuperación de la información. Altavista utiliza plantillas con segmentos de frases en lenguaje natural las cuales relacionan el recurso informacional con el tema de búsqueda, lo que facilita que el usuario pueda utilizar la opción más eficaz para la realización de la búsqueda, esto no siempre sucede ya que el sistema de búsqueda trata de listar los elementos de búsqueda de acuerdo a la información que se consulta frecuentemente, lo que hace que esta búsqueda sea marcada por lo que otros hacen.

El metabuscador Ask Jeeves, que utiliza en su sistema de detección de texto procedimientos de ingeniería del conocimiento, modela la forma en que los humanos realizan sus búsquedas en la Web en una base de conocimiento operada mediante técnicas de inteligencia artificial, lo que facilita utilizar todos estos saberes en la búsqueda de información.

#### 3.6.3.3. Plantillas de Metadatos

El auge de la documentación electrónica unido a las dificultades que acarrea esta para el procesamiento de la información, al ser más costosa, ha conllevado la formulación de diversos formatos de presentación individual, autoformateo (Pinto, 2001). Para lograr dicho procedimiento es vital el conocimiento de los

elementos a incluir en el documento fuente, pues con ellos se podría normalizar en forma de metadatos el proceso extractivo. Estas plantillas se incorporan al formato de extracción para facilitar la recuperación de la información. En el sistema ADAM especializado en Química puede observarse como los documentos son extractados y luego recuperados gracias a la estructura retórica que poseen. ADAN utiliza también un modelo de identificación basado en modelos de contextos. Es por ello que este programa manipula información inherente a secuencias elementales como: en conclusión.... los resultados indican...., etc. (Pinto, 2001). ADAM está compuesto por los elementos siguientes:

- Diccionario: Llamado lista de control de palabras (WLC)
- Un conjunto de reglas que especifican las funciones para cada entrada en la lista.

Este método no resuelve los problemas de coherencia que se solucionan con los aportes hechos por (Mathis et al., 1973) quiénes someten las frases elegidas a diversas transformaciones, de modo que si estas se consideraban adecuadas para la inclusión en el resumen y necesitaban un antecedente, entonces se examinaban las frases anteriores a este para determinar si se incluían también. De no ser así, la frase seleccionada se reescribiría de modo que tuviera sentido por sí misma y si esto no era factible se eliminaba. En esta línea de trabajo también se establece Proteus.

Del estudio de estas técnicas se puede se concluye que las técnicas de nivel medio que más se conocen en la actualidad se encuentra el reconocimiento de cadenas léxicas, entendida como el reconocimiento de relaciones léxico-semánticas (Barzilay and Elhadad, 1999, Silber and McCoy, 2000). Este tipo de extracción de texto se realiza en cuatro fases: primeramente se segmenta el texto original, seguidamente se construyen las cadenas léxicas, se identifican las cadenas léxicas y para finalizar se extraen las oraciones significativas.

En esta línea de tratamiento textual aparece el trabajo de Fuentes (Fuentes et al., 2004), el cual se resume automáticamente documentos escritos en catalán (noticias de prensa) a partir de la extracción de fragmentos, aprovechando las posibilidades que brinda la cohesión textual a través de la detección de cadenas léxicas, ya sean de correferencia o de nominalización oracional. En Fuentes y Rodríguez (Fuentes and Rodríguez, 2002) aparece la aproximación teórica que devino en este sistema.

Es conocido que los resúmenes automáticos poseen ciertos inconvenientes de presentación como: incoherencia, desbalance textual y falta de cohesión. Esta se da sobre todo por la deficiencia que se observa cuando una oración de un resumen por extracción contiene una anáfora cuyo referente está en una oración anterior que el sistema no ha seleccionado como relevante. Esta situación no ha pasado inadvertida para algunos investigadores, en cuyos trabajos sobre las técnicas de nivel medio analizan la correferencia y las referencias anafóricas que se establecen en los segmentos oracionales para mejorar la calidad de los resultados pragmáticos en los textos extraídos.

Específicamente Boguraev y Kennedy (Boguraev and Kennedy, 1997) desarrollan un sistema de resolución de anáfora sustentado en el cálculo de la relevancia en consonancia con factores claves entre los que se encuentran: parámetros gramaticales, sintácticos y contextuales, determinando las palabras previamente y seleccionándolas como candidatos.

Recientemente en las aportaciones de Orasan y Borovets (Orasan and Borovets., 2007) se ha visto como la solución de los problemas de la anáfora pronominal beneficia los sistemas de resumen automático, si se utiliza para la detección de términos. El método se basa en el supuesto de que la relevancia de una oración en el texto está condicionada a las palabras que se encuentran en ella. Esta lectura de este fenómeno trae también inconvenientes, pues considerar a las palabras de forma individual eliminaría el análisis de las relaciones anafóricas. Por ello Oransan (2007) desarrolla un módulo de resolución de anáfora pronominal en una herramienta sumista que se basa en la

identificación de términos, donde se observa si se obtiene mayor calidad en el procesamiento textual, experimento que le ha brindado buenos resultados.

Otra técnica de nivel medio es la que se sustenta en la Máxima Relevancia Marginal (MRM), la misma parte de la selección de la oración núcleo del texto, calculando luego la relevancia marginal de las restantes oraciones usando una fórmula de MRM. Finalmente se selecciona la oración de mayor relevancia y la une al resumen. El proceso se detiene cuando la longitud es la deseada y se reordenan las oraciones (Goldstein, 1999).

#### 3.6.3.4.- Software y Sistemas hechos sobre estructura profunda

Las técnicas de nivel profundo son ideas que sustentan la necesidad de resumir los textos con mayor nivel lingüístico y explotar la estructura discursiva del texto, un reclamo de los investigadores de muchas ramas del conocimiento como: lingüistas, sociólogos, psicólogos y expertos en información.

**SUMMONS (SUMMarizing Online NewS articles).**

Realiza sumarios a partir de bases de datos convencionales o bases de conocimiento. Extracta los contenidos de fuentes informativas de diversos orígenes, pero centradas en un mismo tema. Este sistema opera con el concurso de plantillas generadas en los procesos extractivos. Las plantillas se introducen en un planificador de contenidos que decide la información a incluir en el sumario, utilizando un grupo de operadores (McKeown and Radev, 1995). Esto propicia que el sistema mejore y distribuya la información que inicialmente esté incompleta y el contraste con conocimientos mejores y más actualizados, por tanto el sistema trabaja a niveles de reglas de inferencia. Las operaciones fundamentales que se realizan en este sistema son las siguientes:

- **Contradicción**: Contradicción entre dos fuentes.
- **Adición:** Integración de los hechos adicionales.
- **Refinamiento:** Trata con más precisión la información que se ha añadido posteriormente a los datos generales.

- **Concordancia:** Facilita informes de las fuentes.
- **Superagrupamiento:** Combina información de fuentes con información incompleta, lo que origina sumarios más detallados.
- **No Información**: Busca informaciones que a veces no están en la fuente, que sin embargo, son inferibles.
- **Tendencia**: Avizora la coincidencia entre los resúmenes.

El generador lingüístico de SUMMONS opera mediante un selector de frases que facilita la detección de las palabras, así como la estructura sintáctica del sumario. También posee una ontología que sirve de intermediaria en los procesos de búsqueda de información.

**FRUMP**

Es un sistema de resúmenes que facilita el desarrollo de resúmenes de artículos cortos. Está construido mediante la simulación de los procesos humanos, gracias a ello FRUMP es capaz de interpretar información basándose en datos y expectativas derivadas de un modelo cognitivo dado. La pragmática y la semántica de este sistema se articulan a través de bases de conocimiento que le dan la capacidad de predecir acontecimientos diversos. Esto hace que FRUMP -sobre la base de nuevos análisis- pueda reformular situaciones y predecirlas.

**SUSY**

Permite la comprensión y generación de sumarios de textos científicos o especializados. Opera en comunicación con los usuarios a los que entrega los sumarios. Al ser un sistema interactivo o de diálogo facilita que el usuario sea capaz de describir la estructura del texto que se debe sumarizar y qué estructura considera válida para el sumario. Posee un analizador sintáctico que opera en tres niveles básicos:

- **Comprensión de Frase**: Formula desde un texto de entrada en lenguaje natural una representación proposicional para la cual hay una especificación disponible.

- **Análisis de la Estructura Textual**: La representación lineal básica esta dentro de la representación lineal ampliada.
- **Orden de los elementos**: Los elementos se ordenan jerárquicamente, de acuerdo con su importancia, dando lugar a una red jerárquica proposicional que representa un conocimiento superior al texto de entrada.

Otro sistemas que se incluyen en esta categoría son **SCISOR** descrito por Rau (Rau, 1987), SCISOR (System for Conceptual Information Summarization Organization and Retrieval), el cual está diseñado para realizar resúmenes mediante la aplicación de herramientas de Ingeniería del conocimiento. Logra extraer información sobre la actividad empresarial (transacciones de empresas y responsabilidades corporativas). Procesa noticias cortas a partir de diversas publicaciones especializadas del mundo empresarial entre las que se encuentra el *Wall Street Journal*. Mediante el lenguaje Kodiac (Lenguaje de Ingeniería del conocimiento) este sistema logra almacenar diversos modelos memorísticos representados mediante estructuras conceptuales. SCISOR, a diferencia de los sistemas comerciales facilita la recuperación conceptual de la información. Permite la formulación de interrogantes, ofreciendo a los usuarios respuestas de poca complejidad en lenguaje natural. El sistema posee una base de conocimiento especializada con una amplitud (gran cantidad de términos y frases del léxico que describe) considerable.

SCISOR cuenta de los siguientes Módulos:

- **Selección y Filtración de la Información:** Permite la selección y filtración de noticias en lenguaje natural sobre empresas y corporaciones a partir de documentos en línea, herramientas de comunicación y modelos de adquisición automática.
- **Análisis Sintáctico Parcial:** Este procedimiento de SCISOR permite establecer dos niveles de análisis sintáctico: el análisis gramatical total o exhaustivo mediante el cual se obtienen interpretaciones de los

significados y el análisis gramatical seccionado que facilita la intercepción de palabras procedentes del texto y la supresión de las erróneas. Gracias a TRUMP[19] que posee un potente analizador sintáctico puede ejecutar las operaciones de filtrado, análisis gramatical y sintáctico de los corpus textuales que deben ser resumidos (Pinto, 2001). En TRUMP existe un intérprete de significados que genera conceptualizaciones y sus respectivas relaciones tomando como referencia las estructuras sintácticas para proceder a su interpretación. Un conjunto de herramientas facilitan el control del léxico de entrada. Con todos estos procesos se genera una base de conocimientos o lexicón computacional.

- Aplicación del Generador King (Knowledge Intensive Generator). Es un sistema generador de frases desarrollado y explotado a partir de diversas estructuras semánticas. A partir del concurso de diversos métodos heurísticos facilita el desarrollo de sumarios de diversos tamaños, que se constituyen con cadenas de oraciones debidamente categorizadas en lenguaje natural. El trabajo con King cuenta de tres fases: **mapeo** (mapping) mediante esta operación se lleva a cabo la recuperación y asociación, lo que facilita la relación de los conceptos con determinadas unidades estructurales lingüísticas, **la selección de plantillas** que permite seleccionar estructuras gramaticales seleccionadas de la base de conocimiento y confrontarlas con las estructuras procedentes del mapeo. Las restricciones por su parte facilitan el desarrollo y determinación de determinadas plantillas. Los documentos que ingresan al sistema SCISOR se integran a otros procesos: el procesamiento textual y la aplicación de estrategias cognitivas y por otro lado, la indización global y la sumarización de los textos sometidos a proceso.

[19] Paquete de software de comprensión de texto

### 3.6.4.- Métodos gráficos y relacionales

Salton (Salton, 1997) diseñó un modelo de recuperación conocido como espacio vectorial, en el que las unidades informativas se representan por grupos o vectores de términos que permitan obtener clúster de documentos y pasajes. Una herramienta que caracteriza esta forma sumista es el sistema SMART que propicia descomponer y estructurar los documentos en segmentos de longitud: secciones, grupos de frases adyacentes o frases sueltas. Salton plantea un esquema de relaciones entre los textos y su retórica para generar mapas relacionados apoyándose en la teoría de grafos que muestran las similitudes de los textos y pasajes que han adquirido determinado valor.

Esta propuesta obliga al trabajo con tres fases:

- Identificación del tema textual.
- Selección de las palabras clave.
- Especificación del umbral de obtención
- Extracción de las oraciones con más peso donde se encuentran las palabras clave.
- Obtención del extracto.

Mediante el uso de las ideas sobre generación de enlaces hipertextuales (elemento distintivo de este procedimiento) se pueden crear enlaces intradocumentales entre los párrafos y/o frases de un texto. Este método que basa sus acciones en las estructuras hipertextuales puede resultar eficaz en el tratamiento de la estructura HTML. El principio que utiliza esta herramienta es la detección de marcadores discursivos, usando básicamente conjunciones y adverbios. Cabe insistir en que la aplicación de este tipo de método en algunos sistemas pertenece aún al más estricto campo de la investigación (Pinto, 2001).

Los tratamientos superficiales son los que se utilizan habitualmente en los productos comerciales. Abordan el texto simplemente como cadenas de caracteres — suele darse el caso de que tratan con letras pero también pueden

ser números o cualquier otro tipo de símbolo— y realizan con ellos operaciones de cálculo para seleccionar algunos como representación del documento. No se considera que este método pueda ser eficaz en el tratamiento de resúmenes en entornos de elevado nivel semántico.

Arco (Arco, 2007) refiere que otro método clásico muy utilizado en el tratamiento automático de resúmenes consiste en la selección de las palabras de mayor extensión en el documento, basado esencialmente en n-gramas y bi-gramas. Este método opera seleccionando como resumen del texto las cadenas de caracteres que comienzan en mayúscula y terminan en punto. Mediante este procedimiento es posible resumir de forma aislada e incoherente los textos.

Otro grupo de métodos se basa en la posición de los términos en el texto, en el párrafo, en la profundidad de la clasificación de secciones, etc. Se trata de seleccionar los fragmentos de texto que ocupan las posiciones que prometen ser más relevantes. Es una propuesta muy eficaz para resumir noticias periodísticas. Genera resúmenes mediante la elección del primer párrafo de la noticia. Este procedimiento tiene dudosa aplicación en artículos científicos ya que la relevancia puede estar en cualquier parte del texto a no ser que éste sea un texto estructurado.

### 3.6.5.- Métodos basados en Entidades

Los métodos basados en entidades parten de técnicas que permiten reconocer unidades lingüísticas entre el núcleo de la cadena de caracteres alfanuméricos y el texto. Para ello es necesario el concurso de analizadores morfológicos y desambiguadores léxicos, ya que una misma cadena puede ser nombre o verbo o pertenecer a otra categoría (Pinto, 2001). Es importante también detectar qué cadenas superficialmente diferentes pertenecen a un mismo concepto. Para ello es necesario disponer de programas lematizadores que propicien llegar a análisis más sofisticados de tipo sintáctico o semántico.

Estos métodos pueden ser capaces de tomar las entidades analizadas con anterioridad y detectar disímiles relaciones establecidas, entre las que se encuentran las siguientes:

- Recurrencia.
- lema.
- semánticas.

Esto permite edificar una representación de la conectividad del texto, en términos de teoría de grafos, de manera que se pueda determinar qué partes del discurso son esencialmente relevantes para el trabajo, por tanto define el valor que pueden tener los segmentos textuales para determinado resumen. Este tipo de construcción textual facilita una extracción mucho menos ambigua. Para implementar un método de construcción de resúmenes con este tipo de herramienta es preciso tener reconocidas diversas herramientas, sobre todo contar con bases de conocimiento léxico como EuroWordNet y reconocedores de entidades y sistemas de resolución de referencias anafóricas.

Estos medios de construcción de resúmenos que establecen sus análisis en las distribuciones del texto digital resultan efectivos en el procesamiento de los documentos que poseen ordenación HTML.

Las aportaciones más recientes en el tema de estructura profunda se centran en la Rhetorical Structure Theory (RST) de Mann y Thompson (Mann and Thompson, 1990) con el fin de determinar la estructura discursiva de los textos para construir resumen (Ono et al., 1994, Marcu, 1999, Marcu, 1997). Los sistemas de resumen que se encuentran en esta vertiente están sustentados en la estructura retórica del texto y sus relaciones discursivas, teniendo en cuenta las posiciones nucleares. Daniel Marcu (Marcu, 2000) basa su teoría (RST) en la división del texto en unidades discursivas de pequeño tamaño y en las relaciones que se establecen entre ellas. Marcu (Marcu, 2000) desarrolla un sistema, en que al crear de forma automática la estructura discursiva del texto utiliza un algoritmo para asignar peso y orden a cada elemento del discurso.

Según D´Cunha (D'Cunha, 2006) (cuanto más alto esté el elemento en la estructura, más peso tendrá, y a la inversa), seleccionando para el resumen los elementos estructurales con mayor peso, suprimiendo aquellos que posean el peso más bajo.

No es esta la única aproximación que explota la estructura discursiva, indiscutiblemente existen otras posturas relativas a este asunto, en donde se aborda el tema desde otra dimensión. Teufel y Moens (2002) han realizado un método para la construcción de artículos especializados en lingüística computacional a partir de la retórica de los textos. Los autores anteriormente mencionados proponen un algoritmo que se apropia de la estructura retórica del texto en el cual aparecen siete categorías (aim, textual, own, background, contrast, basis, other) que facilitan la clasificación de los textos. Finalmente el algoritmo referido es capaz de seleccionar los contenidos que se deben incluir en el resumen a partir de diversas categorías.

Existen diferencias entre los trabajos de Teufel y Moens (Teufel and Moens, 2002) y los de RST (Marcu, 1999, Ono et al., 1994). Teufel y Moens no tienen en cuenta la jerarquía para trabajar con la retórica del texto. También Teufel y Moens (2002) intentan capturar la retórica textual en consonancia con el texto y su significado, sin embargo RST se basa en la posición estructural de cada elemento del texto.

El autor considera pertinente la revisión del trabajo de Alonso (Alonso and Fuentes, 2002), en el cual se declara un modelo de estructura discursiva cuya utilidad en el resumen automático ha sido probada, debido a la exploración por parte de la autora de las bondades de la estructura discursiva el texto (aprovecha la puntuación, la retórica, las marcas de discurso), lo que le facilita la representación del discurso y la detección de unidades de relevancia dentro del documento, o sea la autora se centra en un análisis de nivel profundo.

### 3.6.6.- Métodos híbridos para la confección de Resúmenes de Texto

Estos procedimientos se basan en un principio híbrido y operan con el concurso de la relación hombre sistema. Estos métodos proponen la confección de extractos antes del acto de resumir (Pinto, 2001). Los métodos híbridos son los que asisten al resumidor en el proceso de creación de pre-resúmenes. Dentro de esta vertiente sumista también se encuentran las siguientes variantes:

- Pre-resumen: Resúmenes construidos en forma de banco de trabajo. Es un texto que va siendo redefinido hasta que se considera que su calidad informativa es significativa para ser un resumen.
- Resumen a la carta: Permite al resumidor segmentar de forma individual que estructura textual usará para la construcción verbal.
- Resumen asistido por ordenador: Dentro de esta tipología se encuentra el software denominado TEXNET: Sistema informático experimental y su principal objetivo es facilitar al resumidor el acto sumista de un modo híbrido, combinando acciones netamente humanas con las de los ordenadores

Dentro de este apartado últimamente han aparecido método específicos capaces de combinar técnicas lingüísticas, sin embargo su uso no está muy desarrollado, debido al coste de las actividades previas al desarrollo de la técnica.

Hoy en día se han propagado sistemas capaces de combinar diversas técnicas de superficie (cue phrases) con el objeto de elevar la relevancia de los fragmentos del texto original, como ocurre, por ejemplo, en los trabajos de Hovy y Lin (Lin and Hovy, 1998) y Mani y Bloedorn (Mani and Bloerdon, 1998a). En los trabajos de Mateo (Mateo et al., 2003) se mezclan técnicas de análisis superficial con técnicas de nivel medio entre las que se encuentran la resolución de anáforas y el uso de conectores discursivos como mejora de la coherencia del texto, combinando estas técnicas superficiales con otras basadas en la detección de anáforas, para mejorar la coherencia del resumen resultante. El

sistema cuenta con un sub-módulo que permite volver a procesar el resumen (reprocesamiento lingüístico) y detectar en él 750 conectores discursivos en el comienzo de la oración. Según los autores, la presencia en el resumen de una oración con uno de los conectores es vital para detectar la oración que le antecede, de lo contrario el resumen que se logra es desbalanzado e ilegible. Como apunta D´cunha (D'Cunha, 2006) en líneas generales, este submódulo, al encontrarse con una oración comenzada por un conector, constata si la oración anterior forma parte del resumen (en este caso mantiene el conector) o no (en este caso, lo elimina). Sin embargo, existen otros sistemas cuya operatoria se basa en la mezcla de técnicas lingüísticas de alta complejidad. El trabajo de Alonso y Fuentes (Alonso and Fuentes, 2002) integra las propiedades de cohesión de un texto, mediante relaciones de coherencia utilizando para ello cadenas léxicas y la estructura retórica, esto se obtiene mediante marcadores de discurso. Es importante destacar que estos autores ya habían declarado algunos trabajos relativos a la coherencia.

Hemos dejado para el final, la valiosa contribución de Aretoulaki (Aretoulaki, 1996, Aretoulaki, 1997), autor del resumidor denominado COSY-MATS capaz de seleccionar oraciones, dicho sistema se basa en rasgos de contenido de tipo pragmático y retórico, determinados mediante criterios lingüísticos superficiales. Estos rasgos son determinados gracias a la existencia de un corpus de 160 artículos de prensa y 170 artículos científicos con sus resúmenes. Dicho corpus se procesa a nivel superficial y a nivel profundo, para ello se toma como sustento diversas teorías como la de los "Communicating Agents", así como la Teoría de los Actos de Habla defendida por Austin y Searle (Austin, 1962, Searle, 1975) además de la Rhetorical Structure Theory (Mann y Thompson 1988), o postulados teóricos centrados en la cohesión y la coherencia textual, entre los que se encuentra la Lingüística Sistémico-Funcional (Halliday and Hasan, 1976). Específicamente se ha trabajado sobre la base de la relación comunicativa Elaboración, de la Rhetorical Structure Theory, o con los conceptos de Function Word o Common Content Word Pools como apunta

(D'Cunha, 2006). El resultado de este análisis facilitó la identificación de 87 rasgos que sirven de base a la selección textual, agrupados en tres grupos: pragmático, intermedio y superficial. El sistema finaliza sus operaciones estableciendo correlaciones entre los grupos anteriormente declarados y los diversos puntos de marca a nivel superficial, indicadores de pistas para detectar los contenidos esenciales del resumen. En otros sistemas de corte híbrido se usan ontologías y mapas de conocimiento que sirven como herramienta de desambiguación y como sistema de recuperación de la información en la Web.

## *3.7. Técnicas descritas para la confección de Resúmenes de Texto*

La variedad de técnicas y procedimientos informáticos utilizados para el procesamiento del resumen es tal que en esta investigación se ha decidido presentar solo aquellas que se usan con más frecuencia en los procesos de extracción de textos

### 3.7.1.- Resumen Monodocumento

La forma más conocida para construir resúmenes extractos de documentos simples es teniendo un programa selector de fragmentos relevantes desde el documento y luego combinarlos en un extracto (Jackson and Moulinier, 2002).

En Jackson y Moulinier (2002) y en Fukumoto (Fukumoto, 2003) se reportan varias técnicas de resumen extracto de un único documento. Algunas de estas técnicas son: resumen por selección de oraciones, resumen por selección de párrafos, resumen basado en discurso y resumen basado en co-referencia. A continuación describiremos cada una de estas estrategias y ejemplificaremos con sistemas reportados en la literatura.

#### 3.7.1.1.- Resumen por selección de oraciones

Un modo común de abordar el desarrollo de un resumen y muy difícil de transformarlo en una tarea más sencilla que permita realizar la mayor parte del trabajo. Uno de los problemas que acarrea este método se da al generar resúmenes extractos, teniendo en cuenta que la oración es frecuentemente (no siempre) seleccionada como la unidad para construir resúmenes. Los problemas

que reporta este proceder pueden mitigarse si a partir de la selección de oraciones reducimos la generación del resumen expresándolo con métodos de clasificación de las oraciones, de modo que las oraciones relevantes o no sean identificadas para formar parte del extracto. Tener una clasificación de oraciones permite incluir o excluir oraciones según la longitud deseada del resumen (Jackson and Moulinier, 2002).

La generación de resúmenes a partir de la selección de oraciones tiene sus ventajas y desventajas en dependencia del texto a resumir y de la longitud del resumen deseado. Es ventajoso utilizar esta estrategia cuando deseamos resumir noticias o textos no extremadamente largos. Un problema de esta estrategia es que los resúmenes resultantes son a menudo, desunidos y no se leen bien. La clasificación de las oraciones, además, tienen obviamente sus ventajas y desventajas respecto a usar unidades mayores o unidades pequeñas.

#### 3.7.1.2.- Resumen por selección de párrafos

Como mencionamos anteriormente, problemas al considerar las oraciones como unidad básica al resumir son la desunión y la falta de coherencia, es decir, no se leen bien los resúmenes. La utilización de componentes básicos más grandes puede contribuir a la obtención de un resumen más coherente. Así, un enfoque puede ser asumir en un texto un número de párrafos que se pueden considerar mejores o relevantes, y que pueden describir el contenido del texto en su totalidad. Esto es particularmente efectivo para ciertos tipos de documentos, donde uno de los primeros párrafos, típicamente el primero, proporciona una descripción coherente de lo que sigue (Jackson and Moulinier, 2002).

La selección de párrafos ha sido bien estudiada, aunque no tiene tanta popularidad como la selección de oraciones. Ella es ventajosa si el resumen requerido es relativamente largo, o si el material es tal que el aspecto principal de un documento es probablemente contenido en un párrafo simple (Jackson and Moulinier, 2002).

### 3.7.1.3.- Resumen basado en discurso

Ya se ha dicho que resumir documentos teniendo en cuenta las oraciones o párrafos como unidades tiene sus ventajas y desventajas. Según Van Dijk (Van Dijk, 1978) el discurso es la forma de representar los modos de comunicación de determinado dominio mediante el análisis de los medios de representación de conocimientos. La estrategia de resumen basado en discurso propone determinar inicialmente la forma típica de discurso del documento a resumir, es decir, modelar inicialmente la estructura del documento que será resumido. Para ello, es necesario inicialmente dividir el documento en unidades de discurso coherentes. Los bloques de texto obtenidos deben reflejar los subtópicos contenidos en el texto. Jackson y Moulinier (2002) plantean que para realizar la segmentación es necesario analizar léxicamente el texto, utilizar una medida de recuperación de la información en aras de determinar la extensión de los bloques, e incorporar un diccionario y un algoritmo de desambiguación léxica estadística.

Esta estrategia tiene gran utilidad porque los tipos diferentes de documentos tienen variaciones de su estructura, y por tanto, es muy útil identificar inicialmente la presencia o la ausencia de segmentos claves, para posteriormente realizar el extracto a partir de las unidades básicas detectadas. Una desventaja de esta estrategia es que generalmente funciona correctamente para un tipo particular de documentos y en un contexto específico.

### 3.7.1.4.- Resumen basado en co-referencia

Co-referencia es un fenómeno lingüístico donde dos o más expresiones pueden representar o indicar la misma entidad. Así como otros fenómenos lingüísticos, la co-referencia puede admitir la ambigüedad. El concepto básico de co-referencia está presente cuando dos expresiones lingüísticas, tales como ‘Bill Gates’ y ‘The Chairman of Microsoft’, se refieren ambas a la misma entidad.

Los métodos anteriormente mencionados están concebidos para obtener resúmenes genéricos, mientras que los métodos basados en co-referencia están

más enfocados a generar el resumen teniendo en cuenta las preguntas de los usuarios y extrayendo el contenido relevante a partir del entorno de la consulta realizada por el usuario. Las asociaciones que existen entre términos que tienen co-referencia pueden ser usadas para clasificar y seleccionar oraciones del documento al ser incorporadas en un resumen. Los métodos que siguen esta estrategia son capaces de generar resúmenes que son casi tan efectivos como el texto completo, ayudando al usuario a determinar la relevancia (Baldwin et al., 2000). La desventaja de esta estrategia es que se requiere realizar un preprocesamiento más costoso del documento. Se necesita entonces etiquetarlo, trabajar con tesauros, etc.

Los métodos analizados no constituyen herramientas especiales para la confección de resúmenes hipermediales. Su campo de acción no sólo se circunscribe a la hipermedia, si no a todo tipo de documento que circula en INTERNET. Se limitan a un único campo de acción: el texto verbalizado. Si bien no ignoran las cualidades esenciales del texto hipermedial, (forma de estructura lineal-reticulada) no actúan sobre otras estructuras textuales como: imagen y sonido, por tanto aunque muchos teóricos de la Ciencia de la Computación (Arco, 2008) insisten en llamarlos herramientas para resumen hipermedial no pueden considerarse como tales. Ellos constituyen medios para la confección de resúmenes textuales operantes con palabras representadas por métodos que se estructuran a través del análisis de los elementos del párrafo.

### *3.8.- Complejidades del tratamiento del Resumen Multidocumento*

Si el resumen de un documento simple se dificulta, resumir múltiples documentos presenta aún más problemas. Sin embargo, el éxito en este empeño ofrecerá una utilidad real a muchos investigadores y “trabajadores de ciencia”, por permitirles procesar colecciones enteras de documentos con menos esfuerzo que en la actualidad (Jackson, 2001).

En Stein (Stein et al., 2000) se plantea que el resumen de documentos simples es sólo una de las tareas críticas necesarias para formar un resumen completo de múltiples documentos. Ejemplos de esto se mencionan a continuación:

- Identificar los temas de importancia en la colección de documentos.
- Seleccionar el resumen de un documento simple representativo por cada uno de los temas.
- Organizar los resúmenes representativos para formar el resumen final de múltiples documentos.

Resumir múltiples documentos es un subtópico que se aborda mucho en la actualidad dentro de las investigaciones sobre la obtención de resúmenes de documentos. Inicialmente, se pensó que se podían aplicar las técnicas usadas para resumir un único documento para obtener el resumen de múltiples documentos considerando la colección de todos los textos como un único documento. Aunque muchas de las técnicas usadas en el resumen de un único documento pueden ser usadas también en el resumen de múltiples documentos, existen al menos tres diferencias significativas (Goldstein, 2000):

- El grado de redundancia en la información contenida en un grupo de documentos relacionados es mucho mayor que el grado de redundancia en un documento.
- Un grupo de documentos puede contener una dimensión temporal, por ejemplo cuando nos referimos a reportes noticiosos.
- El tamaño del resumen con respecto al tamaño del documento será típicamente mucho menor para colecciones de documentos que para un único documento.

El problema de la co-referencia en los resúmenes presenta retos mayores al resumir múltiples documentos que al resumir un único documento.

En James, y Gupta (James and Gupta, 2001) se describen varias formas de crear resúmenes por extracto de múltiples documentos:

- Resumir secciones comunes de los documentos: Encontrar las partes importantes y relevantes que la colección de documentos tiene en común (su intersección) y usarla como un resumen.
- Resumir secciones comunes y secciones únicas de los documentos: Encontrar las partes importantes y relevantes que la colección de documentos tiene en común y las partes relevantes que son únicas y usarlas como un resumen.
- Resumir el documento centro (más representativo): Crear el resumen de un único documento a partir del documento centro o más representativo de la colección.
- Resumir el documento centro (más representativo) más el resto de los documentos de la colección: Crear el resumen de un único documento a partir del documento centro o más representativo de la colección y agregar alguna representación del resto de los documentos (pasajes o extracción de palabras claves) para proveer un cubrimiento total de la colección de documentos.
- Resumir el documento más reciente más el resto de los documentos de la colección: Crear el resumen del documento que tiene información más reciente y adicionar alguna representación del resto de los documentos para proveer un cubrimiento de la colección de documentos.
- Resumir secciones comunes y secciones únicas de documentos teniendo en cuenta el factor tiempo: Encontrar las partes relevantes e importantes que la colección de documentos tiene en común y las partes relevantes que son únicas. Tener en cuenta la secuencia de tiempo de la información extraída en la generación del resumen.

### 3.8.1.- Principales algoritmos de agrupamiento de Documentos

Los algoritmos de agrupamiento han sido técnicas que han contribuido al desarrollo de sistemas para la extracción de textos, pues permiten la

clasificación y el agrupamiento de documentos. Son operaciones matemáticas que reflejan las posibilidades exactas de llevar a cabo con fiabilidad una extracción y una clasificación (Ver Figura 10).

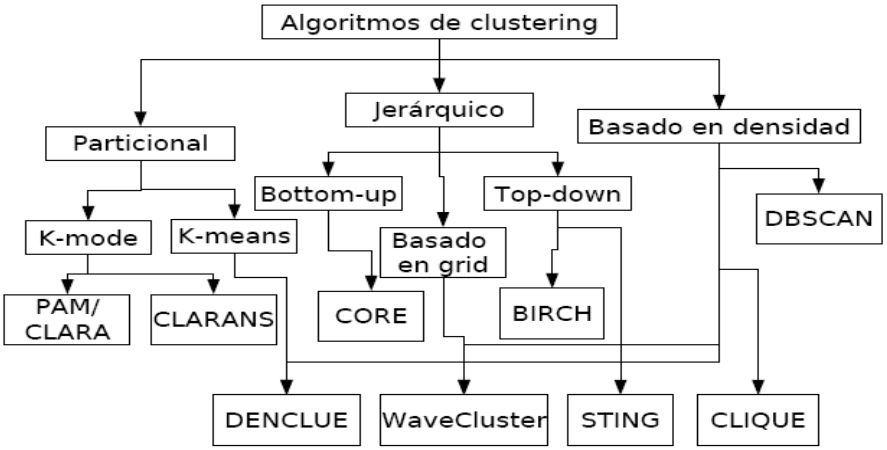



Figura 10 Tipología de los algoritmos de clustering

Estos procederes matemáticos han dado lugar al desarrollo de algoritmos borrosos, entre los que se encuentran: FUZZY SkWIC, SKWIC y MLRul que permiten obtener y agrupar datos cuya estructura no se encuentra en un conjunto definido. Es importante destacar que en este trabajo sólo se describen algoritmos cuya relevancia incide en el desarrollo de esta temática. Para ello se ha decidido clasificar los algoritmos en las clasificaciones siguientes: basados en grafos, algoritmos de agrupamiento basados en densidad, algoritmos para la construcción de resúmenes supervisados y algoritmos no supervisados para la construcción de resúmenes.

### 3.8.1.1.- Algoritmos Basados en Grafos

Existen varios algoritmos que se sustentan a través de grafos. Uno de los más usados es el *k*-medias (*k*-means) que tiene una elevada complejidad temporal, según Jain y Dubes (Jain and Dubes, 1988), Kaufman y Rousseeuw (Kaufman and Rousseeuw, 1990), Ying y Karypis, (Ying and Karypis, 2002) y Xiong, et al. (Xiong et al., 2006). Esta complejidad reside en la capacidad del algoritmo de

trabajar con varios grupos sin que se establezca un límite para el agrupamiento. Si se analiza su rendimiento se puede constatar que dicho algoritmo tiene limitaciones de operación. Su eficacia depende de un elemento esencial: la convexidad en los grupos que deben construirse o lograse.

Esta convexidad es la capacidad que tienen los elementos o los conjuntos que deben agruparse, permitiendo que dentro de ellos las conexiones se establezcan en línea recta. Es importante destacar que para lograr altos niveles de eficacia, este sistema necesita entre sus condiciones operativas que el número de grupos a obtener sea determinado de antemano, por tal motivo, requiere conocer algunas especificidades del dominio que se representa, pues utiliza elementos de la partición inicial. K-means es un algoritmo base, pues ha servido de modelo para la construcción de varios algoritmos como el *x*medias (*x*-means) que permite estimar de forma eficaz el número de grupos de un conjunto dado, el conjunto *k*-medias (batch *k*-means) y el *k*-medias incremental (incremental *k*-means), este último cuenta con una nueva versión que lo hace más efectivo y que se denomina medias (means). Esta versión mejora el rendimiento del algoritmo según los trabajos de Berry (Berry, 2004) en los cuales se describen estudios de similaridad que representan valores más promisorios que los de otros algoritmos.

En este terreno también se destacan como algoritmos PAM (*Partitioning Around Medoids)* de Kaufman y Rousseeuw (Kaufman and Rousseeuw, 1990), y sus versiones CLARA (*Clustering LARge) Application)* y CLARANS (*Clustering LARge ApplicatioNS).* Estas dos versiones del algoritmo son muy efectivas y su diferencia está en que el resultado de la clasificación con CLARA produce grupos cuya homogeneidad es menor que los obtenidos con CLARANS. Han y Kamber (Han and Kamber, 2001) han desarrollado variantes sobre este algoritmo al igual que Agarwal y Mustafa (Aggarwal and Mustafa, 2004 ).

Estos algoritmos pretenden enmendar las deficiencias operativas y de rendimiento asociadas al *k*-medias, es decir pretenden obtener grupos sin tener conocimiento del dominio, sin embargo, aunque son más eficaces que k-medias,

el problema persiste. Además estos algoritmos en su mayoría son costosos, ya que sus gastos de operación (lento) son muy elevados. Estos algoritmos son eficientes si los datos tienen determinadas condiciones como baja dimensionalidad, los conjuntos tienen una marcada definición y gozan de alta densidad. Otras notaciones que descuellan sobre *k*-medias en lo referente a resultados pragmáticos son aquellos que operan bajo análisis discriminatorio de conjuntos entre los que se encuentran los descritos por Torre y Kanade (Torre and Kanade, 2006) y Bolelli (Bolelli et al., 2007), este último utiliza una mezcla operacional con SVM (*Support Vector Machines)* según Bordes (Bordes et al., 2005).

Otro algoritmo que mejora el *k*-medias se denomina *primero más lejano* descrito en (Hochbaum and Shmoys, 1985) y en Ramírez y Montes de Oca, (Ramírez and Montes de Oca, 2004). En su operatoria se realizan las siguientes acciones:

Selección aleatoria de los núcleos centrales de grupos que se van a obtener. (Ver Anexo 5)

Cálculo de la distancia de cada instancia al centroide de más proximidad a esta.

Este proceso es repetido hasta que el número de grupo sea mayor que un umbral especificado o sea se intenta lograr que la instancia que se encuentra más próxima sea aceptada como el centroide de un grupo, por tanto este algoritmo tiene como inconveniente la necesidad de hacer reducciones en la dimensionalidad para obtener resultados esperados.

Este algoritmo es utilizado en el agrupamiento de documentos según Liu (Liu et al., 2006). El algoritmo EM (*Expectation-Maximization)* (1998) y su mejora FREM (*Fast and Robust Expectation Maximization)* son también herramientas de este corte, que según Ordóñez y Omiecinski (Ordoñez and Omiecinski, 2002) asigna a cada instancia una distribución de probabilidad, lo que hace que pueda determinarse el rango en que se encuentran los elementos que pertenecen a cada clúster. Aunque los referidos métodos de agrupación manipulan datos de alta dimensionalidad, realizan un refinamiento muy costoso.

La representación y el procesamiento de la información tienen nexos indisolubles con la teoría de grafos, pues se han erigido no sólo como herramientas de probada validez para formular modelos diversos de agrupamiento, sino que también es un medio esencial para demostrar matemáticamente elementos y asociaciones que ocurren a nivel sensorial. Es el sustento esencial para el desarrollo de instrumentos y herramientas de análisis cuyo valor pragmático facilita la realización de diversas tareas y operaciones.

Otra de las posibilidades de la teoría de grafos es su capacidad de servir de fundamento teórico metodológico para la construcción de algoritmos de agrupamiento y sus formas de validación. Los grafos admiten la formulación de instrumentos para construir análisis de grupos. La literatura científica describe muchos instrumentos basados en grafos. Entre ellos se encuentra Estrella o (Star) descrito por Aslam, et al. (Aslam et al., 1998), que ha sido modificado por (Gil-García et al., 2003) y por (Medina and Pérez, 2007) facilitando la clasificación efectiva de todos los elementos necesarios por el investigador.

La característica principal de Estrella Extendido (Extended Star, ES) es su operatividad y efectividad con respecto a otras variantes como Estrella de (Gil-García, 2003), pues permite obtener menor número de grupos y por tanto son más consistentes los conglomerados. Estrella permite más particiones de grupos, sin embargo desde el punto de vista de agrupamiento los grupos tienden a ser menos representativos y menos homogéneos. Otras variantes de estos algoritmos son Estrella Generalizada (Generalized Star; GStar) de Medina y Pérez (2007) y Estrella Condensada (Condensed Star; ACONS) de Medina y Pérez (2007) en los que se proponen nuevas visiones del concepto central de estrella y propician la obtención de menor cantidad de grupos o clúster.

Su principal característica reside en la necesidad de especificar la cantidad de conjuntos a obtener, sin embargo, el resultado final depende del umbral o del rango determinado antes de hacer el agrupamiento. En este campo aparece el algoritmo MLDS (*Multilingual Document Clustering)* descrito en (Wei et al., 2008), que realiza el agrupamiento de los textos a partir de colecciones de

documentos multilingües, extrayendo del corpus textual aquellas frecuencias de términos a partir del un idioma especificado. Este algoritmo tiene mucha efectividad y aplicación en la construcción de diccionarios de equivalencias y lexicones computacionales.

### 3.8.1.2.- Algoritmos de agrupamiento basados en Densidad

Los algoritmos de mayor relevancia basados en densidad son:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Kriegel and Pfeifle, 2005).

DENCLUE (*DENsity-based CLUstErin)* (Hinneburg y Keim, 1998).

OPTICS (Ordering Points To Identify the Clustering Structure) (Ankerst et al., 1999).

Estas formas algorítmicas, tienen una complejidad $O(n \log n)$. Según (Ankerst et al., 1999) es denominado comb sort o algoritmo de ordenamiento, que indica la velocidad del algoritmo y su efectividad está en la localización de parámetros determinados. Este proceder específicamente supera el algoritmo de agrupamiento de burbuja, por lo que su rendimiento no es efectivo cuando se trabaja con datos de alta dimensionalidad, además su rendimiento depende altamente de diversos elementos iniciales que deben estar especificados en el conjunto que debe agruparse o particionarse.

Por su parte OPTICS de Ankerst, et al. (1999) y la propuesta de Kriegel y Pfeifle (2005), se erigen como variantes mejoradas del DBSCAN. Otros algoritmos basados en densidad son los desarrollados por Ruiz-Shulcloper, Alba-Cabrera y Sánchez Díaz, (Ruiz-Shulcloper et al., 2000), Qian, et al. (Quian et al., 2004) y Dourisboure, et al. (Dourisboure et al., 2007). Otros autores como Falkowski, et al. (Falkowski et al., 2006) y Hu y Wu (Hu and Wu, 2007) también desarrollan notaciones basadas en la densidad local de los nodos mediante las cualidades de los grafos scale-free como Kalisky, et al. (Kalisky et al., 2006), Fortunato (Fortunato et al., 2006) y Stumpf (Stumpf et al., 2005).

### 3.8.1.3.- Algoritmos basados en Celdas

Los algoritmos esenciales basados en celdas son: STING (*STatistical INformation Gris)* descrito por Wang, (Wang et al., 1997) y Wave Cluster Sheikholeslami (Sheikholeslami et al., 2000), desarrollado en (1998) y CLIQUE (*CLustering In Quest)* de Aggarwal y Yu (Aggarwal and Yu, 2005). Los referidos algoritmos poseen un alto nivel de escalabilidad y su complejidad es *O*(*n*), sin embargo no son efectivos si se intenta agrupar información o datos de alta dimensionalidad, pues su operatoria está sujeta a la existencia de patrones sustentados sobre la estructura geométrica de objetos espaciales, por tanto estos algoritmos no están formulados sobre medidas de distancia.

Agrupar elementos utilizando la densidad como modelo facilita el descubrimiento de conglomerados diversos, sin embargo si se utiliza agrupamiento basado en celdas se obtiene rapidez en la operación. AGRID (*Advanced Grid-based Iso-Density line)*, CLONE (*Clustering with Low-Order Neighbors*) reportado por Arco (Arco, 2008) y GARDEN (*Balanced Iterative Reducing and Clustering using Hierarchies* descrito en Orlandic, et al. (Orlandic et al., 2005) constituyen casos que mezclan los enfoques anteriormente descritos. Es indiscutible el nivel que han tomado los algoritmos jerárquicos en la literatura especializada de la especialidad según Jonyer., et al. (Jonyer et al., 2002), a pesar de tener altos niveles de complejidad.

BIRCH de Zhang, et al. (Zhang et al., 1996) se erige como una variación dentro de este algoritmo y posee complejidad lineal. Sin embargo, es ineficiente en el desarrollo de grupos, ya que indiscutiblemente necesita nuevos elementos de entrada que estén vinculados con la longitud de los grupos que se pretende obtener. Según Allen (Allen, 2005) su efectividad operacional depende de las características de los datos de entrada, si estos poseen alta dimensionalidad el algoritmo trabajará con un bajo rendimiento.

CURE, por su parte, permite la construcción de grupos de variadas formas y tamaños (Allen, 2005). Posee altos niveles de densidad, siendo su complejidad,

*O*(*n*2log *n*), esto significa que requiere de optimizaciones para lograr un rendimiento efectivo. Tanto en BIRCH como en CURE se aprecia un desarrollo correcto de la detección de puntos que no se encuentran en un grupo o rango determinado. BIRCH posee menor complejidad algorítmica, y debido a esto, su capacidad de agrupamiento es muy baja. El algoritmo PDDP (*Principal Direction Divisive Partitioning)* descrito por Boley (Boley, 1988) y su versión mejorada sPDDP (*Spherical Principal Directions Divisive Partitioning)* son los ejemplos más citados en el terreno de las técnicas jerárquicas divisivas Berry (Berry, 2004). Las salidas de ambos son utilizadas en los procesos de entrada de los algoritmos de medias. Gran importancia revierten las variantes concatenadas en el desarrollo de los algoritmos, estas sirven para la formulación de herramientas basadas en la lógica difusa.

Se destacan en esta dimensión los trabajos desarrollados en la Universidad Central de las Villas, Cuba por el Grupo de Inteligencia Artificial dirigidos por Bello (Bello et al., 2006) el cual desarrolla, a partir del método Estrella Extendida de Gil-García (Gil-García et al., 2003) un proceder que facilita la construcción de algoritmos borrosos del tipo SKWIC (*Simultaneous Keyword Identification and Clustering of text documents)*. Con este sistema algorítmico es posible obtener un agrupamiento de mejor calidad y, además, no requiere poseer un conocimiento previo del dominio. En autores como Cheng (Cheng et al., 2006) aparece un nuevo procedimiento que mezcla dos estrategias: la divisiva y la aglomerativa.

Esto reafirma el uso combinado de los métodos anteriormente referidos en la representación de la información y el agrupamiento. Algunos algoritmos se apegan a estrategias meramente divisivas como los siguientes:

Construcción de un árbol de expansión mínimo desarrollado por Zahn (2010).

Enlace simple Gower y Ross (Gower and Ross, 1969), Gotlieb y Kumar (Gotlieb and Kumar, 1968).

Enlace completo de Backer y Hubert (Backer and Hubert, 1976).

En los trabajos de Gil-García, et al. (2006) se describen propuestas para el trabajo con algoritmos jerárquicos aglomerativos basados en grafos. Además, existen dos modalidades algorítmicas que se nutren de las representaciones gráficas de datos son STIRR *(Sieving Through Iterated Relational Reinforcement)* de Gibson, et al. (Gibson et al., 1998), la cual utiliza métodos espectrales; una variación aglomerativa cuya complejidad es $O(nk+n \log n+k2\log k)$ que facilita la construcción de grafos de *k*-vecinos más cercanos, particionando y combinando los grupos por su inter-conectividad y cercanía (Karypis, et al., (1999) y Allen, (Allen, 2005) varias tipologías de datos provenientes de otras aplicaciones del agrupamiento pueden ser modelados como grafos de doble partición según Gao, et al. (Gao et al., 2005), Deodhar y Ghosh, (Deodhar and Ghosh, 2007) condición que facilita el desarrollo de aplicaciones en campos como la bioinformática y las telecomunicaciones.

Existen otras propuestas que se desarrollan a partir de la explotación de las propiedades estructurales del grafo y que han sido desarrolladas por Radicchi, et al, (Radicchi et al., 2004). Esta aportación es especialmente valiosa si se trabaja con gráficos muy densos con clara definición entre los grupos fuertes y débiles, considerando el grado de las conexiones internas en los grupos y las conexiones externas hacia otros grupos. Este método basa sus acciones en el coeficiente de agrupamiento local, y su complejidad temporal o coste de ejecución se expresa como $O(m4/n2)$, lo que indica que se necesita un elevado trabajo de optimización del algoritmo para lograr que su ejecución sea segura. En esta misma vertiente se ubica el algoritmo SCAN *(Simultaneos Multidocumets identifications)* de Xu, et al. (2007), cuya complejidad $O(m)$ utiliza la vecindad de los nodos como teoría de agrupamiento, es decir bajo la suposición de los nodos y los vecinos y la cercanía entre ambos.

El algoritmo SMTIN (*Simultaneos Mutidocumets identifications)* facilita el minado de datos en el espacio virtual, sin embargo necesita que se le determine un umbral de distancia, lo que impide la obtención de múltiples resoluciones (Epter and Krishnamoorthy, 1999). Estos problemas operativos se solucionan si se

apela a la modificación desarrollada por Epter y Krishnamoorthy, (1999), los cuales logran establecer una variante que no requiere del umbral de distancia y reporta resultados múltiples en el agrupamiento. Otras propuestas basadas en propiedades netamente estructurales se presentan en Cortes, et al. (Cortes et al., 2001), Aggarwal y Yu (Aggarwal and Yu, 2005) y Wasserman y Faust (Wasserman and Faust, 1994).

#### *3.8.1.3.1.- Algunas aplicaciones de los Algoritmos de Clasificación*

El desarrollo de algoritmos de agrupamiento posee grandes aplicaciones, sin embargo es difícil valorar en el desarrollo de una aplicación la posición preponderante de un algoritmo u otro. Es evidente que en el uso de las técnicas clasificatorias se mezclan algoritmos y variantes de los mismos para lograr las aplicaciones deseadas. Esto confirma que no hay un algoritmo único e indispensable. Estos se recombinan y facilitan el desarrollo de nuevas aplicaciones. En la mayoría de los casos sólo se aplican a proyectos muy puntuales y en otros exclusivamente sirven para comprobar o para ilustrar la efectividad de otros. Las aplicaciones o las soluciones pragmáticas aún son restringidas, lo que evidencia que en terreno de la clasificación queda mucho por decir. Aquí se intentará describir algunas aplicaciones puntuales de algunos de los algoritmos que se han descrito en el acápite anterior.

En los trabajos de Montejo-Ráez, (Montejo-Ráez, 2005) se observa cómo se obtienen clases a partir de documentos con estructura XML y aplicadas al procesamiento de registros bibliográficos, utilizando para ello SVM (Vector Space Model) y PLAUM. Estos algoritmos posibilitan la obtención de clases depuradas o de clases sin ruido. Para comprobar su efectividad se utiliza como medida la precisión, f-measures y recall, esto permitió validar que los algoritmos basados en análisis léxico son superiores a los que no hacen análisis de dominio.

En algoritmo K-means ha servido con medidor de la precisión de otros algoritmos, donde se han utilizado algoritmos genéticos para la descripción y

desarrollo de la clasificación automática, ya que las mismas permiten construir mutaciones y permutaciones. Otra aplicación del K-means esta descrita en la tesis de Marinelli, (Marinelli, 2002) en la que se utilizan el algoritmo anteriormente referido para identificación de hábitos de uso de sitios Web utilizando redes neuronales. Este algoritmo ha servido para desarrollar el basamento teórico de la clusterización. En este experimento también se utiliza el algoritmo CLARA (Clustering LARge Applications) de Kaufman y Rousseeuw (Kaufman and Rousseeuw, 1990) para la limpieza y obtención de los datos del experimento.

En el terreno de la clasificación también sobresalen los trabajos de Montalvo (Montalvo, 2006). Mediante el uso del lexicón computacional *EuroWordNet 1.0* y mediante el empleo de un método de desambiguación automática, los autores describen diversas funciones para el pesado de los rasgos (TF, TF-IDF y WIDF) para obtener clustering multilingües utilizando para ello la librería CLUTO, y más concretamente Direct, una versión particional muy semejante a las que se usan para agrupar documentos de diversos idiomas. El algoritmo es especialmente efectivo cuando se trabaja con las funciones TF y TF-IDF.

En el caso de OPTICS, Ankerst (Ankerst et al., 1999) se describe que su desarrollo permite obtener resultados favorables en el tratamiento visual de la estructura de los agrupamientos, esto permite dar visibilidad a las zonas esenciales del Clúster y poder usar el agrupamiento para visualizar la clasificación, algo que puede ser utilizado en sistemas de búsqueda y recuperación de la información.

Existen en la literatura diversos casos de aplicación de algoritmos de agrupamiento donde se utilizan medidas de similaridad y formas de agrupamiento. Tal es el caso de (Leiva et al., 2009) que emplea el Modelo de espacio vectorial que, a su vez, se basa en lo que se ha denominado bolsa de palabras interceptadas mediante vectores, donde las palabras frecuentes son utilizadas para la obtención de grupos. A través de este algoritmo, desarrollado a partir de vectores, se logra la localización de documentos y la plasmación de

relaciones entre ellos a través de la similaridad o distancia. Esta forma de agrupamiento es muy efectiva si se desconoce el dominio con que se está tratando. VSM también ha sido utilizado en Corpus Miner para facilitar el agrupamiento empleando técnicas de Clustering (Arco, 2007). Este sistema también aporta resúmenes extractos cuya calidad es baja. Los algoritmos que utiliza en su método de agrupamiento son la estructura concatenada son SKWIC -Extended Star - Fuzzy SKWIK.

Del análisis de los agrupamientos utilizando precisión y exhaustividad se ha obtenido que los algoritmos concatenados anteriormente mencionados son más efectivos que las variantes no concatenadas descritas en otros procesos similares. Dentro de estas herramientas sobresale el sistema GarLucene que permite la realización de búsquedas sobre los documentos indexados, especificar la localización, creación y actualización de los índices, selección de algoritmos de agrupamiento y el desarrollo de medidas de validación a utilizar, lo que la convierte en un instrumento eficaz en la solución de los problemas de agrupamiento, además esta herramienta permite la programación de indexaciones, gathers, etc.

Gran cantidad de productos informáticos se han desarrollado para construir agrupamientos, sin embargo su efectividad se lastra por la poca variedad de las aplicaciones que, en su mayoría, se orientan a la clasificación de cuestiones muy específicas.

#### 3.8.1.4.- Algoritmos de Construcción de Resúmenes supervisados

Una dimensión más detallada de la construcción de resúmenes puede observarse al estudiar los algoritmos de resúmenes supervisados. Aunque este tipo de algoritmo necesita de una colección de entrenamiento constituida por pares texto/fuente-resumen. Los resúmenes que se desarrollan a partir de este tipo de algoritmo son esencialmente científicos, debido a que las colecciones de entrenamiento tienen una retórica específica casi siempre. Algunos de los algoritmos más citados en la literatura internacional se describen a continuación:

Algoritmo de Edmundson (1969): Es el algoritmo que encabeza el desarrollo de los métodos de extracción de textos. En este algoritmo cada segmento del texto es representado a partir de técnicas de vectorización. Un vector de cuatro componentes es asociado a los rasgos (palabras-pista, palabras-clave, palabras – título y localización de la sentencia) según Pons (Pons, 2006). La misma autora sostiene que "a cada sentencia se le asigna una puntuación definida como $W(s)=\alpha C(s)+\beta K(s)+\gamma L(s)+\delta T(s)$ donde $C(s)$, $K(s)$, $T(s)$ y $L(s)$ denotan, respectivamente, los valores de los rasgos *palabras-pistas*, *palabras-claves*, *palabras-título* y *localización* de *s*, y α, β, γ y δ son sus pesos asociados". Al ser un algoritmo basado en ponderación, se obtienen pesos de las sentencias a medida que aparecen las palabras en una colección supervisada. Es importante destacar que mediante este algoritmo se logra identificar con facilidad los elementos que se encuentran al inicio y al final de los corpus textuales, algo que es muy efectivo para el proceso extractivo, ya que las sentencias de mayor carga informativa se encuentran generalmente al inicio y al final del texto fuente. Si se analiza un texto estructurado retóricamente puede apreciarse que las estructuras retóricas de mayor carga están en determinado segmento del texto entre los que se encuentran: (la introducción y las conclusiones). Muchos especialistas critican la linealidad del tratamiento de las sentencias descritas en este algoritmo (Pons, 2006), La misma autora opina que, sin embargo, es un algoritmo muy usado en el desarrollo de extractos, pues muchas versiones que existen sobre él incluyen sus rasgos. La fórmula para su cálculo es la siguiente:

$$N_1 + \frac{\sum_{i=1}^{N_2} i_{s_1} i_{s_2}}{\sqrt{\sum_{i=1}^{N_2} i_{s_1}{}^2 \sum_{i=1}^{N_2} i_{s_2}{}^2}}$$

donde $P(sE)$ denota la probabilidad, a priori, de que una sentencia de un texto fuente sea incluida en su extracto, y $k\{1,\ldots,n\}$ los valores $P(Rk(s)=vk|sE)$ y $P(Rk(s)=vk)$ son constantes que denotan, respectivamente, la probabilidad de

que el rasgo *Rk* de una sentencia que pertenece al extracto sea igual a la constante *vk* y la probabilidad de que el rasgo *Rk* de una sentencia sea igual a *vk*. Con respecto a este algoritmo es importante destacar que es efectivo para el desarrollo de sistemas para detectar anáforas y determinar los elementos cohesivos del texto.

Algoritmo de Kupiec: Este algoritmo está basado en las sentencias de un texto fuente representadas mediante un vector de valores. Es un algoritmo basado en rasgos y en su desarrollo según (Pons, 2006) se tiene en cuenta la longitud de la sentencia, pues si esta es mayor que un umbral determinado, la localización de la sentencia se establece dentro de los 10 primeros o los últimos párrafos del texto. En similitud con el algoritmo de Edmundson éste clasifica las sentencias de mayor puntuación como pertenecientes al extracto. Es importante destacar que este algoritmo ha servido como referente para el desarrollo de otros algoritmos basados en colecciones de entrenamiento. Otros algoritmos sobre este tema han sido descritos por (Lin and Hovy, 1998)y Mani y Bloedon (Mani and Bloerdon, 1998a, Mani and Bloerdon, 1998b).

Tanto en las aportaciones de Edmunson como en las de Kupiec se utilizan árboles de decisión. Según Pons (2006) “En los algoritmos supervisados de construcción de extractos las operaciones de la fase de análisis (análisis lexicográfico del texto fuente, segmentación del mismo en sentencias, y la obtención de la representación de éstas) y el cómputo del puntaje asociado a las sentencias se pueden efectuar en un tiempo proporcional a $O(|X|)$, donde $|X|$ denota la longitud del texto fuente”. Este algoritmo facilita la construcción de extractos de calidad cuyas medidas de evaluación son suficientemente efectivas, pero además, desde el punto de vista pragmático, es capaz de generar resúmenes de elevada calidad, con el inconveniente de ser efectivo sólo en un dominio lingüístico determinado. La fórmula de este algoritmo es la siguiente:

$$P(s \in E \mid R_1(s) = v_1 \ldots R_n(s) = v_n) = P(s \in E) \frac{\prod_{i=1}^{n} P(R_i(s) = v_i \mid s \in E)}{\prod_{i=1}^{n} P(R_i(s) = v_i)}$$

Según Pons (2006) *P* (*sE*) identifica la probabilidad a priori de que una sentencia de un texto fuente sea incluida en su extracto, y *k*{1,…,*n*} en los valores *P*(*Rk*(*s*)=*vk*|*sE*) y *P*(*Rk*(*s*)=*vk*) son constantes que denotan, respectivamente, la probabilidad de que el rasgo *Rk* de una sentencia que pertenece al extracto sea igual a la constante *vk* y la probabilidad de que el rasgo *Rk* de una sentencia sea igual a *vk*.

### 3.8.1.5.- Algoritmos de construcción de resúmenes no supervisados

Los algoritmos no supervisados se desarrollan en dos vertientes. La primera variante se formula a partir de la ponderación de cada uno de los elementos del texto por separado. Para ello se tiene en cuenta las condiciones estadísticas y lingüístico-semánticas del texto, lo que permitirá detectar los elementos de mayor peso, al igual que el procedimiento descrito por Edmundson. En trabajos de Goldstein, (Goldstein, 1999, Goldstein, 2000) pueden verse investigaciones relacionadas con este tipo de algoritmo, que se emplea con mucha profusión en sistemas comerciales. Otros algoritmos que se identifican con esta vertiente utilizan el discurso textual para extraer el texto resultante. Los más importantes en esta tipología son los siguientes:

Algoritmo de Barzilay basado en la cohesión léxica (Barzilay and Elhadad, 1999). “Este algoritmo para cada cadena léxica fuerte del texto fuente, selecciona la primera sentencia del texto que contiene un miembro representativo de la cadena y la clasifica como perteneciente al extracto” (Pons, 2006). Cuando se calculan las cadenas léxicas se hace el cálculo a partir de las relaciones entre las palabras tomando como patrón una base de conocimiento de carácter ontológico como WordNet o EuroWordNet. Sin embargo, estas bases de datos son generales y se hacen inoperantes para elementos

especiales del texto. En la praxis muchos investigadores comienzan a desarrollar sus propias bases de conocimiento léxico, pues la forma en que se construyen estas ontologías toman como referentes elementos meramente gramaticales y no los contextuales, por tanto si bien han servido como medio de evaluación de algunos presupuestos teóricos, es indiscutible su poca capacidad para brindar un producto terminado para la realidad comunicacional del usuario que es más rica semánticamente, incluso que los modelos que intentan calcar el pensamiento humano.

Algoritmo de Nomoto y Matsumoto (Nomoto and Matsumoto, 2001). El algoritmo realiza una partición del conjunto de las sentencias del texto fuente a través de un procedimiento de agrupamiento. Esta notación denota el extracto como el conjunto formado por la sentencia más importante de cada subconjunto. La elección de tales sentencias se realiza teniendo en cuenta la frecuencia de sus términos en el documento. Presenta nexos con la teoría de grafos pues sus nodos representan los elementos del léxico y las aristas las relaciones lexicales a través de las que se da la cohesión textual. Según Pons, (2006) este algoritmo presenta complejidad de tipo cuadrático cuando se analiza contra el texto fuente, es decir, ha de ser optimizado para que su velocidad y su coste sean los deseados.

Muy pocos son los algoritmos utilizados para el tratamiento de los textos de forma léxico-semántica, este es uno de los problemas que aún revisa la Ciencia de la Computación, mientras que otras disciplinas demandan cada vez más representaciones textuales basadas en modelos socio-cognitivos y sociopragmáticos. “Los algoritmos que se basan en la cohesión léxica se erigen como un grafo cuyos nodos sirven de la representación de palabras, frases y grupos de frases representadas por sentencias y aristas. La complejidad temporal de estos algoritmos es cuadrática con respecto a la longitud del texto fuente” (Pons, 2006). La literatura reporta la existencia de muchos algoritmos y sistemas de construcción de extractos basados en la cohesión de un texto.

Ejemplos de ellos son: ERSS *System de* (Bergler, 2004)y *K.U. Leuben Summarization System.*

Este análisis evidencia que no existe superioridad de ningún algoritmo con respecto a otro, su eficiencia reside en lo que se quiera solucionar y sus características. Indiscutiblemente el análisis del dominio facilita determinar el tipo de agrupamiento que se va a utilizar y por ende supondrá mejores resultados pragmáticos. Es indudable que existen muchos algoritmos de agrupamiento, pero muy pocas propuestas toman en consideración las ventajas de las estructuras inherentes al lenguaje. En los trabajos de (Ferrer and Solé, 2001) se demuestra la validez del léxico en la intermediación textual y el agrupamiento.

En varios trabajos se constata que el lenguaje existe en una red, cuyo valor semántico y cognitivo tiene implicaciones sociopragmáticas más que algorítmicas y matemáticas, sin embargo a pesar de que se reconoce el desarrollo de estos elementos, son escasas las ocasiones en que la literatura de la especialidad declara su uso en agrupamiento de texto. La mayor parte de los algoritmos que se utilizan en minería de texto son herramientas monodimensionales ancladas netamente en presupuestos matemáticos, esto evidencia que a la Ciencia de la Computación no le interesa el dominio lingüístico y socio-pragmático, pues es más fácil construir un método de agrupamiento que se “adapte” a cualquier lenguaje y que sea instrumentado con facilidad. No obstante comienzan a aparecer algunos razonamientos matemáticos que se desarrollan a partir del estudio del discurso.

En el caso de los algoritmos que se usan particularmente en el desarrollo de los extractos, se puede señalar como limitante la falta de cohesión y balance textual, lo que hace que los extractos tengan que ser revisados mediante un sistema de revisión de textos. A esto se suma la falta de estilo en el extracto generado, esto desencadena bajas valorizaciones en el proceso de evaluación, pues hay muchas diferencias del extracto resultante respecto a la colección de entrenamiento.

Los nuevos algoritmos para el desarrollo de resúmenes deben estar enrutados al desarrollo de resúmenes de mayor calidad, es decir con mejor balance textual y cohesivo. Estas herramientas deben prestar atención especial al desarrollo de estrategias para modelizar el tratamiento de la anáfora pronominal en Redes Semánticas y Ontologías.

### *3.8.2. – Los Métodos de Desambiguación Automática*

Se conoce como desambiguación automática al conjunto de técnicas que se encargan de darle un sentido a los vocablos en determinado contexto. Muchos son los métodos desarrollados para estos fines entre los que se encuentran dos vertientes esenciales: los supervisados y los no supervisados. La distinción entre estos dos tipos de métodos se hace difícil, sin embargo es aceptado en toda la comunidad investigadora que los métodos supervisados se sustentan de reglas de clasificación o modelos estadísticos para DSA generalmente usando corpus o léxicos etiquetados manualmente, en tanto los métodos no supervisados se sustentan en sentidos o clases mediante elementos no etiquetados. Actualmente coexisten métodos que usan tanto ejemplos no etiquetados como no etiquetados, ejemplo de ello son los lexicones y diccionarios electrónicos, los cuales son denominados métodos mínimamente supervisados. En esta investigación solo se exponen los métodos que sirven de postura para el trabajo presente y futuro de TEXMINER, es por ello que se mencionan los métodos mixtos de desambiguación como una alternativa de trabajo y no se explican ya que no interesan en este tipo de investigación restricta a un dominio específico.

#### 3.8.2.1.- Métodos Supervisados

Entre los métodos supervisados se encuentra el aprendizaje automático (AA, en inglés, Machine Learning, ML), cuyo objetivo es formular algoritmos que faciliten la obtención de la descripción coherente de un concepto subyacente dentro de un conjunto de observaciones. Esta descripción debe ser coherente con el conjunto de observaciones y debe permitir predecir futuras observaciones del mismo problema (Márquez, 2002).

Las diversas técnicas de AA se pueden clasificar según distintos parámetros:

- el conocimiento adquirido y representado tiene carácter netamente simbólico y en última instancia puede ser incluso subsimbólico;
- el aprendizaje no posee dualidad y puede ser supervisado o no supervisado;
- el desarrollo de cada técnica tiene un sustento indispensable en los modelos estocásticos conocidos también de razonamiento inductivo.

Lo enunciado en el acápite anterior indica que estos métodos dificultan la identificación de clases debido a la gran cantidad de combinaciones que se establecen entre los mismos, aspecto que Márquez (Márquez, 2002), especifica en sus cinco familias de métodos de aprendizaje automático:

- **Métodos basados en aprendizaje estadístico:** Se encargan de aprender todos los modelos estocásticos que sean capaces de modelar los datos, logrando determinar el proceso que ha sido capaz de brindar mayor calidad en datos observados. En esta corriente aparecen las redes bayesianas, las cadenas ocultas de Markov y los métodos basados en el Principio de Máxima Entropía.

  **Métodos tradicionales de la Inteligencia Artificial:** En esta clasificación se insertan todos los métodos que incluyen aprendizaje simbólico entre los que se encuentran: árboles de decisión, listas de decisión, aprendizaje de reglas, inducción de programas lógicos o razonamiento basado en casos. Una de las bondades de estos métodos estriba en que el conocimiento representado es muy fácil de interpretar y manipular. Es importante también describir dentro de estos métodos a los que denominamos métodos de aprendizaje subsimbólico entre los que se encuentran las redes neuronales, las que evidencian resultados similares a los obtenidos con métodos enteramente simbólicos,independientemente que el resultado o sea la representación del conocimiento que se genera con estos instrumentos no se interprete igual por los humanos.

**Métodos de la Teoría del Aprendizaje Computacional** (Computational Learning Theory): Los principales algoritmos generados en esta área, cuya influencia en el PLN ha sido evidente son: Winnow, SnoW y Support Vector Machines entre otros.

Se encarga del estudio integral del aprendizaje, más que un método parece ser una subdisciplina. Sus principales valores están en la detección de instrumentos para identificar el tamaño óptimo de las muestras de aprendizaje, el espacio de hipótesis, su complejidad y el grado calidad con que sucede o con que se logra la aproximación del concepto objetivo, entre otras más.

- **Métodos semi-supervisados:** Su función se centra en mejorar las deficiencias que acarrean los métodos supervisados. Los métodos semi-supervisados intentan generar modelos de elevada confiabilidad mediante el uso de colecciones pequeñas, combinadas en algunos casos con grandes y medianas colecciones de ejemplos no supervisados. En esta clasificación aparecen los algoritmos que a continuación se declaran: algritmos de bootstrapping, Expectación-Maximización (Manning and Schütze, 1999) el co-training o las Transductive Support Vector Machines.

### *3.8.2.- Tipos de DSA*

#### 3.8.2. 1.- DSA Supervisada

Tal y como se expuso al inicio estamos frente a un tipo de DSA donde los ejemplos deben ser etiquetados. En esta DSA siempre se aplican algoritmos de aprendizaje para extraer de determinados corpus aquellos atributos que son necesarios para construir una representación de cada uno de los sentidos. Esta forma de representación es aplicable a diversas actividades de la vida cotidiana. Generalmente lo que se representa aquí son las ocurrencias y la co-ocurrencias de palabras, expresadas como lemas o categorías sintácticas. Casuísticamente se trabaja el orden de los vocablos y de las relaciones sintácticas, tratando habitualmente las palabras de contexto como un conjunto desestructurado,

aunque habitualmente las palabras del contexto se tratan como un conjunto sin estructura.

El autor ha decidido presentar algunos métodos que han sido utilizados en la obtención de atributos a partir de ejemplos diversos:

- **Clasificadores bayesianos simples** (naive): Es naive el clasificador más sencillo entre los estocásticos y funciona estimando la probabilidad de desambiguar una palabra X que tiene un sentido particular. La hipótesis que sostiene esta clasificación está en el teorema de Bayes (regla de inversión de dependencia entre eventos) y la simplificación de la independencia de atributos:

$$P(clase\_i|E) = \frac{P(clase\_i) \times P(E|clase\_i)}{P(E)}$$

- **Modelos descomponibles**. Estas formulaciones son una dimensión general del método de Bayes simple (naive Bayes): los rasgos tratados se agrupan en subconjuntos mutuamente dependientes. Es importante destacar que la independencia entre rasgos solo se establece si los subconjuntos son diferentes, no se da en un mismo conjunto. Como se ha declarado anteriormente nos encontramos ante una variante para enfrentar problemas de data sparseness, donde la agrupación de las palabras en clases obliga a disminuir la cantidad de parámetros del modelo.
- **Modelos ocultos de Markov** (Hidden Markov Models, HMM): Es un modelo estadístico que ha sido aplicado con mucho éxito en la estadística para el etiquetado morfológico (POS-tagging). Este modelo es muy similar al funcionamiento de un autómata finito, ya que su función de transición es netamente probabilística. En este modelo los arcos entre los estados están etiquetados con probabilidades y los estados están representando cada una de las variables del modelo en forma de función probabilística.

El modelo funciona encontrando la secuencia de estados, lo que indica un máximo de probabilidad para producir un ejemplo (Márquez, 2002). Este algoritmo ha tenido éxito cuando ha sido aplicado por (Segond et al., 2000).

- **Árboles de decisión**. Método preveniente de la Inteligencia Artificial, siendo un árbol de decisión una secuencia de preguntas, que se sustenta en la segmentación progresiva del conjunto de ejemplos. El árbol se transforma en un camino con cierta univocidad desde la pregunta raíz a la terminal. Los nodos que aparecen en el camino contienen preguntas (una única pregunta) sobre atributos específicos y el arco que enlaza el nodo siguiente presenta un valor para este atributo. La hoja final a la cual lleva el camino ofrece una clase, o sentido, como resultado más probable para ese ejemplo (Márquez, 2002). Este método es deficiente si se quiere utilizar para fragmentar a la manera de la DSA, pues trata gran cantidad de datos.
- **Listas de decisión**, Es una lista ordenada de reglas conjuntivas, las reglas se analizan de forma ordenada sobre el ejemplo, siendo la primera en cumplirse la usada para realizar la clasificación. El sustento del modelo se sostiene en la hipótesis que posibilita discriminar los ejemplos de un dominio a base de reglas individuales de menos complejidad. Con las listas de decisión se logra impedir la segmentación de los datos, resultantes de la aplicación de árboles de decisión en dominios textuales con pocos ejemplos de aprendizaje, muchos atributos distintos, atributos con muchos valores, etc. (Márquez, 2002). Rivest (1987 citado por Yarowsky, 2000). El algoritmo utiliza las secuencias de reglas condición - acción ordenadas de forma cíclica de más segura a menos segura, obligando a que se utilice la más segura de las reglas. En Yarowsky (1994, 1995) se han utilizado estrategias de aprendizaje interpolado con el fin de mejorar los problemas que por ausencia de datos se encuentran en la construcción de las listas.

- **Vectores de contenido** (Schütze, 1992): Estos vectores se basan en el contexto donde ocurre la palabra de forma similar a como se trata un documento dentro del terreno de la Recuperación de la Información, generando vectores en un espacio n-dimensional, donde n es el número de palabras del documento, o contexto para DSA, cada punto del vector contiene una función que identifica la frecuencia de esa palabra en el documento o contexto. El autor es consciente que existen otras medidas para DSA que difieren de las de RI.
- **Aprendizaje basado en la memoria o basado en ejemplos** (memory based learning). Este tipo de aprendizaje se centra en la memorización de los ejemplos sobre los que trabaja para luego almacenarlos sin ningún procesamiento. La clasificación de un nuevo ejemplo implica la obtención en la base de ejemplos el conjunto de los k ejemplos más semejantes al ejemplo que se va a clasificar, los k vecinos más próximos (k-NearestNeighbours, kNN), determinando la clase que más prevalece entre ellos. Esta fase de clasificación implica la comparación del ejemplo que se quiere clasificar con cada uno de los ejemplos guardados y el cálculo de la distancia entre ellos. La distancia utilizada para el trabajo con ejemplos en esta métrica es la distancia Hamming, denominada métrica de solapamiento en la cual la distancia entre dos ejemplos es calculada sumando las distancias entre los valores de cada atributo (simplificando a 1 cuando estamos en presencia de igualdad y a 0 en cuando hay contraste entre los valores), estando cada distancia asociada con un determinado peso (Márquez, 2002).

#### 3.8.2.2.- DSA No Supervisada

Cuando se adquiere conocimiento léxico existen múltiples problemáticas, por ello el objeto de estos métodos es provocar el conjunto de sentidos de una palabra oscura utilizando para ello un grupo de ejemplos sin etiquetar, lo que hace posible que solo sea utilizado para la discriminación de sentidos. Esto facilita la dependencia de los datos etiquetados a nivel de sentido. De esta

manera, se evita la dependencia de datos etiquetados en cada sentido. Los métodos que caracterizan este tipo de DSA son capaces de representar el contexto de ocurrencia de la palabra al igual que los vectores de contenido. En Schütze (1992) se ha observado la aplicación de medidas de similitud para edificar conjuntos usando los vectores de mayor similitud mediante clustering.

### 3.8.2.3.- DSA semi-supervisada (mínimamente supervisada)

Lo que caracteriza este tipo de DSA es la proposición de los métodos anteriormente desarrollados de forma inversa a partir de la modelación de sentidos como vectores de contenido, aquellos ejemplos que no han sido etiquetados se modelan de la misma forma y son agrupados de acuerdo a la similitud. Cada centro geométrico de cada clase es visto como vectores óptimos para definir el sentido, lo que facilita la utilización de vectores con más ajuste a los datos. Proceso denominado bootstrapping.

Con vistas a impedir la ocurrencia de errores de progresión geométrica en las repeticiones sucesivas en los métodos de bootrapping se aplican dos supuestos:

a. un vocablo mantiene su sentido en todo el trayecto de un texto (onesense per discourse), y
b. que un vocablo generalmente pose el mismo sentido en contextos análogos (one sense per collocation).

Es posible en la experimentación obtener vectores de partida en estos métodos mediante ejemplos etiquetados manualmente o quizás con el uso de tesauros y diccionarios. El algoritmo denominado Expectación-Maximización (EM) es un algoritmo clásico de bootstrapping. Su operatividad empieza mediante estimaciones iniciales de los parámetros del modelo, aumentando la probabilidad para generar datos observados (Márquez, 2002).

## *3.9.- Métodos basados en fuentes de conocimiento estructuradas (knowledge-based WSD).*

La caracterización de los sentidos es una de las posibilidades más utilizadas en DSA, lo que se traduce en utilizar información denotada en fuentes de conocimiento léxico generalmente a través de la caracterización general de los sentidos, esto se traduce en usar información explicitada en recursos que existen de antemano y que son estructuras de conocimiento. Los métodos que caracterizan este tipo de métodos son diferentes en cuanto:

a. la utilización de recursos léxicos (DAO monolingües y/o bilingües, thesauros, bases de conocimiento léxico, etc.)
b. el contenido de estos recursos que sirve para la experimentación;
c. el uso de propiedades en el momento de relacionar los léxicos.
d. la vertiente del proceso de DSA que utiliza.

### 3.9.1. - Tipos de DSA Knowledge-based WSD

3.9.1.1.- Basados en Diccionarios

En los trabajos de Lesk (1986) aparece por primera vez la utilización de los DAO como soporte de las estructura de información léxica para la DSA. Todos los enfoques y métodos de desambiguación que se sustentan en diccionarios son amplios en variedad y en complejidad. Generalmente estos tipos de DSA utilizan como solución computacional la asociación de palabras con elevados niveles de polisemia al primer sentido de la entrada del diccionario, pues la lexicografía en sus artículos léxicos usa de forma general en esta posición el término más habitual o más frecuente. Es este método una de las heurísticas tomadas como patrón en la construcción de sistemas de DSA. Las definiciones en los diccionarios se suelen usar para la DSA como bolsas de palabras (bag of words).

#### 3.9.1.1.2.- DSA basada en tesauros y en la información temática de los diccionarios

La desambiguación basada en tesauros explota la caracterización semántica y estructural de los tesauros como el DECS (Descriptores en Ciencias de la Salud)

o como diccionarios temáticos como el Longman's (LDOCE, Procter 1978). Los códigos temáticos (subject codes) son utilizados para etiquetar sentidos especializados, de un dominio. En Walker y Amsler (1986, citados por Yarowsky, 2000) se evidencia como el código temático es la forma más correcta para palabras como bank, a la que se asocian múltiples dominios especializados asociados en un tesauro. La forma en que este modelo es capaz de inferir conocimiento se centra en la capacidad que tienen las categorías semánticas de las palabras de un contexto para indetificar la categoría semántica del contexto en su totalidad.

**3.9.1.1.3.- DSA basada en los códigos temáticos de los diccionarios.**

Otras de las variantes de uso de los diccionarios en DSA está anclada en la utilización de diccionarios con definiciones distintas, donde la herramienta más utilizada es el LDOCE (Longman Dictionary of Contemporary English), que contiene sobre todo, box codes y subject codes para cada sentido. Los primeros enuncian posturas primitivas y semánticas, además emiten códigos de limitaciones o restricciones para los nombres y los adjetivos, y la estructura argumentativa de los verbos. En el segundo caso se utiliza un conglomerado o conjunto de formas primitivas que facilitan la clasificación de los sentidos de los vocablos bajo rúbricas temáticas. Al igual que los Tesauros que utilizan definiciones o clases semánticas en estos métodos se estudian utilizando diversos métodos entre los que se encuentran: las funciones de similitud de Wilks et al. (1993) o las técnicas de clasificación óptima de Guthire y Cowie basadas en temple simulado. Uno de los problemas de esta técnica reside en que a veces es inadecuada para desambiguar dominios específicos debido a la caracterización de las palabras de forma general. En Yarowsky (1992) se explicita el uso de un clasificador en un corpus, solo que su uso está sustentado a la condición de que un sentido se propaga a través de varios temas, y es entonces que el algoritmo tiende a ser ineficiente y la solución que se le ha dado entonces es hacer distinciones de sentido independientes del tema. Para

Stevenson y Wilks (2000), los métodos basados en tesauros trabajan mejor para nombres que para verbos.

**3.9.1.1.4.- DSA con diccionarios bilingües**

Hemos dejado para el final de este tema el uso de diccionarios bilingües. El sustento teórico que acoge a este método es la existencia de diferentes sentidos que una palabra tiene y los diferentes sentidos de esta cuando se traduce. Indiscutiblemente el análisis de las diferentes traducciones de diversos vocablos puede usarse como una pista para acopiar información sobre los sentidos de la palabra de entrada. Si se analizan los algoritmos de Dagan e Itai y el de Manning y Schütze (1999). El algoritmo de Dagan identifica para un nombre determinado, los sentidos de sus ocurrencias en un texto teniendo en cuenta las diversas traducciones de dicho término en otra lengua, en frases donde se declara una función objeto, utilizando para ello un diccionario bilingüe que conecta a ambas lenguas. Básicamente el algoritmo busca en el diccionario las traducciones del léxico resultante y el verbo que lo acompaña. El algoritmo de Dagan e Itai posee elevada complejidad, es capaz de desambiguar si puede tomar una decisión certera y confiable. De existir equilibrio en las frecuencias de coocurrencia de verbos y nombres, y equilibrio entre las traducciones la variante más frecuente implicará un gran error. Es esta la cuestión que hace que en este caso sea necesario tomar esta decisión sí se supera un umbral alto aproximadamente de un 90 %. Más adelante se describirán los algoritmos que trabajan con ontologías y que trabajan con el sistema que se propone en esta tesis. Para finalizar mencionaremos los métodos mixtos para la DSA, los cuales han estado usándose para buscar mejores resultados en la DSA sobre todo para elevar la calidad de la DSA en dominios más generales. Estos métodos son una combinación de los métodos anteriormente expuestos.

### 3.10.- Software para la construcción de resúmenes de Texto

El estudio o conocimiento de los software o herramientas elaboradas para el desarrollo del resumen reviste gran importancia para la elaboración de un método sumista. Matcalf (Matlclaf, 2006) y Moreiro (2004) mencionan varias formas de agrupación de las herramientas. En este caso se ha decidido la que propone Moreiro (2004) que divide las herramientas en dos grupos fundamentales:

- **Productos de investigación.**
  - Mediante extracción de oraciones.
  - De comprensión profunda.
  - Por aproximaciones híbridas.
- **Productos comerciales**

#### 3.10.1. Mediante Extracción de Oraciones

Estos programas se basan en la detección de oraciones en el texto. Los principales son los siguientes:

- Summ – It applet, creado por la Universidad de Surrey el funciona utilzando técnicas de cohesión lexical http://www.mcs.surrey.ac.uk./System/summary/
- SweSum creado en Suecia por el Instituto de Tecnología. Este software genera extractos basados en textos de publicaciones seriadas de carácter científico. Mantiene un estrecho vínculo con el ISI. En esta herramienta los conjuntos oracionales se extraen a partir del las premisas declaradas por Luhn en los años cincuentas del siglo *XX*, es decir la extracción se realiza por ponderación.
- Tex Summarization Projet de la Universidad de Ottawa. Proyecto que está desarrollando técnicas de vanguardia en la definición e identificación de palabras claves. Para lograr estas propuestas apelan

al uso de cálculos estadísticos sobre la superficie del texto en la que determinan aspectos lingüísticos. http://www.site.uotta.ca/tan-ka/ts.html.

- FociSum es también un sistema que opera por detección de oraciones es desarrollado por la Universidad de Columbia y emite resúmenes de corte científico (sobre artículos científicos) empleando herramientas de análisis sintáctico y semántico mediante los cuales es capaz de identificar conceptos. Elabora resúmenes indicativos utilizando plantillas preestablecidas. http://ww.isi.edu/natural-language/pro-jets/SUMMARIST.html.
- ISI Summarist. Es un programa que mediante el sistema de traducción Systran produce extractos de artículos de noticias de prensa. Delimita los temas esenciales de un documento mediante procederes detallados que determinan la posición de los temas y las frases en el documento. También se sirve de la superestructura del discurso. Con las oraciones del discurso se generan resúmenes de elevada coherencia.

### 3.10.2.- De Comprensión Profunda

Utiliza mecanismos semánticos y contextuales para la generación de extractos:

- Trestle: Text Retrieval, Extraction and Summarization for large Enterprises. Es un agente desarrollado por la Universidad de Sheffield. El mismo plantea un uso de artes de avanzada para extraer la información que proviene de la realidad. http://www.dcs.shef.ac.uk/re-search/groups/nlp/trestle.
- Sumons es un sistema que elabora noticias de prensa a partir de documentos múltiples. Se desarrolla por la Universidad de Columbia. Su estrategia extractiva utiliza las plantillas con elementos de semántica predefinida, lo que facilita la producción de textos de elevada calidad. La misma no es totalmente original, pues se nutre de

las plataformas elaboradas en otros proyectos para la selección de contenidos, la constitución de estructuras sintácticas y la concepción de colocaciones de superficie. Es un sistema muy eficaz ya que posee una semántica compleja que se aproxima a las formas de extracción humana. http://www.cs.columbia.edu/-regina/demo4/.

### 3.10.3.- Por Aproximaciones híbridas

Combinan técnicas de extracción con otras más específicas del procesamiento del lenguaje natural:

- MutiGen de la Universidad de Columbia. Utiliza un modelo lingüístico y cognitivo para el procesamiento del lenguaje natural. Elabora extractos mediante la siguiente estrategia: 1) Elimina las frases desconocidas; 2) combina y organiza oraciones reducidas; 3) Hace transformaciones sintácticas. 4) Suple oraciones por paráfrasis; 5) Cambia conceptos 6) Ordena oraciones. http://www.cs.columbia.edu/-hjim /sumDemo/CPS/.

### 3.10.4.- Comerciales basados en extracción

- Datahammer: Resume textos en línea. Trabaja de conjunto con el browser del usuario. La extracción de las oraciones la realiza mediante el algoritmo Microword Tree Triiming. http:// www.glu.com/datahammer/ .
- Text Analyst de Megaputer. Extrae el corpus textual desde la computadora del usuario. Construye una red semántica del documento por lo cual formula un conjunto de procedimientos obtenidos de una red neuronal. La red semántica resultante no depende del dominio ni de sus conocimientos previos. Se le muestra un mapa conceptual al usuario para que seleccione los contenidos. http://www.megaputer.com/html/texanalyst.html.
- Un punto aparte merece la IBM como comercializadora de productos de extracción. Estas herramientas permiten al

usuario seleccionar resúmenes por delimitación de oraciones. Sobresalen dentro de ellas: Internet King of Translation y Lotus Word Pro, ambas en idioma nipón. Los extractos generados se generan sobre la base de los conjuntos oracionales, las relaciones retóricas y la posición de las oraciones dentro del texto. Es importante destacar que la elección de las oraciones de hace de forma estadística.

- In Text Search Enginer Websumm. Es un producto de la corporación Mitre que hace los sumarios sobre documentos individuales o múltiples. Para lograr sus objetivos como producto emplea motores de búsqueda. El extracto se conforma al establecerse nodos entre las preguntas de los usuarios y las oraciones que mejor responden a los términos. http://www.mitre.org/pubs/edge/july_97/.
- In Text producido por de Island Soft es un extractor que opera en la macroestructura textual en consonancia con la presencia en esta de palabras claves. Permite la posibilidad de seleccionar la técnica de extracción más viable para el usuario, lo que hace que este pueda aplicarlos a cualquier documento que tenga guardado en su máquina. http://www.ics.mq.edu.au/-swan/summarization/projets.htm.
- Summary Server. Herramienta sumista producida por LinguisticX que produce extractos con documentos fuera de línea. Se sirve de técnicas estadísticas que utilizan la posición de la oración, su longitud y la presencia en ellas de palabras clave. Esto propicia la obtención de resúmenes de elevada calidad. Se le brinda al usuario la posibilidad de delimitar el tamaño del resumen y la posición de la palabra clave que desee. http://www.inxight.com.

- Extractor de Tetranet Software es capaz de seleccionar las palabras claves de un documento, y a través de ellas encontrar las formas macroestructurales y las oraciones que las contienen. Los procedimientos que sirven de herramientas a este software tienen como parámetro fundamental la longitud de las formas primitivas de las palabras, que luego son cohesionadas mediante procedimientos de (Gen Ex) algorítmicos genéticos. Las frases extraídas se igualan semánticamente. El producto final se muestra mediante un documento TXT y la lista de oraciones es filtrada mediante una heurística. http://www.extractor.iit.nrc.ca.

- Sinope Summarizer realiza de forma automática resúmenes monodocumentales y a la vez conserva las imágenes, el formato y la página que fue extractada. Mediante técnicas lingüísticas de alta generación es capaz de analizar la macroestructura textual y la macroestructura parcial, o sea las partes de la macroestructura. Está disponible de tres formas en Internet: 1. Sinope Personal Edition, (admite a los usuarios generar extractos de las páginas que se muestran en Internet durante su búsqueda. 2. Sinope Server Edition es una herramienta concebida para sistemas de bibliotecas y puede utilizarse para obtener sumarios exclusivamente o para distribuir estos en otros servicios bibliotecarios. 3. Sinope Search Engine Edition opera como un motor de búsqueda a la vez que emite microresúmenes con los resultados de la búsqueda, lo cual facilita la evaluación del servicio. http://www.sinope.nl./en/sinope/index.html.

- Pertinente summarizer. Instrumento creado por ANVAR con el concurso de la Agencia Francesa de Innovación, el cual proporciona la construcción de resúmenes multilingües. Se

adapta a los entornos de XML y J2EE (Java). Opera de forma exclusiva con técnicas lingüísticas que ayudan al usuario en la selección de términos que le prestan ayuda en la búsqueda y en la evaluación del resultado final. Los términos se utilizan también en el acto de resumen. Cada término que aparece en el extracto está acompañado de su frecuencia de aparición y con enlaces hipertextuales al documento original o al mismo resumen. Produce información sobre textos en francés, inglés, español, alemán, portugués e italiano. http://www.pertinence.net/PMwhitepaper.pdf.

- Copernic Summarizer facilita la obtención de resúmenes de cualquier documento o página Web. Extrae toda la información importante mediante complejos algoritmos lingüísticos y estadísticos. Extrae las oraciones más relevantes del texto mediante la aplicación de las teorías macroestructurales. Otros programas relacionados son MS Word Autosumarization de la Microsoft; Discontinued Summarization Products de Apple y Webcompass de Quarterdeck. http://www.copernic.comp.desktop/products/summarizer .

### *3.11.- El Tratamiento del Texto en Cuba*

El desarrollo de las técnicas de tratamiento de texto en nuestro país es aún incipiente. Algunas instituciones de investigación se han dedicado al desarrollo de software para el tratamiento de textos. Entre las que más se destacan: CENATAV[20], Universidad Central de Las Villas y la Universidad de Oriente. Los trabajos no constituyen aún herramientas comerciales, pues su uso está restringido a fases de investigación. La Universidad Central de las Villas desarrolla en el Centro de Estudios de Estudios de Informática adscrito a la

[20] Centro Nacional de Tecnologías de Avanzada.

Facultad de Matemática Física y Computación el software Corpus Miner cuya autora principal es la Dra. Leticia Arco García.

- Corpus Miner ofrece a los investigadores y desarrolladores en el campo de la minería de textos la posibilidad de realizar el agrupamiento de textos la selección de las palabras claves que caracterizan los grupos textuales obtenidos, y la creación de las aproximaciones superiores e inferiores de los grupos textuales. Es un producto que agrupa los documentos utilizando las posibilidades de la teoría de grafos. La agrupación del corpus textual se perfila como una valiosa posibilidad para establecer sistemas de búsqueda y recuperación de información y para la puesta en marcha de servicios de resúmenes. Corpus Miner facilita la selección de las palabras de mayor relevancia resultante de métodos de agrupamiento. La selección de los términos con mayores valores de calidad en el grupo y la selección de los términos a partir de las reglas generadas por el algoritmo ID3 en cualquiera de sus variantes combinadas. http://www.uclv.edu.cu .
- SATEX (Arco, 2008) es una herramienta que permite clasificar un corpus textual sólo en formato txt, obtener las relaciones entre los términos de cada grupo, las palabras claves esenciales y el resumen realizado por el autor. También permite la obtención de métricas de evaluación del agrupamiento.
- Garlucene: (Arco, 2008) Software que permite la indexación, la clasificación de documentos y la obtención de grupos textuales mediante un criterio de agrupamiento jerárquico. Este software se utiliza mucho en el desarrollo de sistemas de búsqueda y recuperación de información debido a sus características como sistema.

El CENATAV también ha generado nuevos algoritmos en función del agrupamiento documental y la búsqueda y recuperación de la información. Entre ellos se encuentran los siguientes:

- Experimentación de Ranking de Documentos de los autores Pérez y Medina (Pérez and Medina, 2007, Medina et al., 2007). Utilizando el Modelo de Distancia Global Asociativa, este considera a los términos relevantes no sólo por su frecuencia de aparición sino también por la relación que mantenga con los restantes términos del documento y la fortaleza de esta relación. A partir de esta consideración se realiza una adecuación del Modelo de Espacio Vectorial y se considera la fuerza global asociativa del término como el rasgo que lo caracteriza y no su frecuencia de aparición. Para las tareas de recuperación en este algoritmo se apela a un nuevo concepto nombrado fuerza de la asociación global, que usa las relaciones semánticas entre los términos en los documentos. Este proceder describe un nuevo modelo de representación de documento nombrado Modelo de Distancia Global Asociativa, con una complejidad nunca mayor que el Modelo de Espacio Vectorial clásico, el cual muestra resultados importantes en la recuperación informativa.
- Compressed Arrays Algorithm for Frequent Patterns (Detección de concurrencias de palabras en colecciones de documentos) Es un algoritmo que permite la aplicación de las RA en áreas de venta y predecir el comportamiento de los clientes. En el caso de Bases de Datos, este agoritmo puede ser usado para construir tesauros estadísticos, en análisis informativo de encuestas y, en tareas de clasificación y agrupamiento. Los autores principales de este procedimiento son (Pérez and Pagola, 2007).
- A Generalized Star Algorithm for Clustering. Útil para la representación de colecciones de documentos, utiliza un grafo de semejanza con umbral. A diferencia de las anteriores variantes que existen de este

algoritmo, este es un nuevo tipo de sub-grafo en forma de estrella llamado ***sub-grafo en forma de estrella generalizada*** que permite obtener grupos de diferentes formas, solucionando las deficiencias de los algoritmos anteriores. Es un mecanismo factible para ser utilizado en sistemas de organización de información como un paso de pre-procesamiento en la representación de información estática. Además puede ser utilizado para tareas de browsing, detección de tópicos, etc.

- Algoritmo para la Segmentación de Documentos por Tópicos. Ventajoso en la Sumarización de Documentos, Segmentación de Noticias, la Recuperación de Información y otras tareas de la Minería de Texto. La mayor dificultad en el proceso de segmentación está en encontrar los límites físicos adecuados de cada temática en el documento. Está dirigido a la segmentación de documentos de múltiples párrafos y particularmente a documentos que explícitamente explican o abordan un tópico especifico. Sus autores son miembros del Cenatav ( http://www.cenatav.co.cu).

### 3.11.1.- Las Investigaciones sobre construcción y extracción de Texto en Cuba

Los esfuerzos para desarrollar métodos de extracción en Cuba no han sido del todo exitosos. A pesar de tener un sistema de ciencia e innovación bien desarrollado, Cuba sólo dispone de tres centros de investigación en los cuales se realiza esta labor. Desde el punto de vista temático los resultados investigativos se agrupan en dos disciplinas esenciales: La Ciencia de la Computación y la Ciencia de la Información. En el terreno de la Ciencia de la Computación la mayoría de las investigaciones son aplicaciones matemáticas para trabajar con sistemas que operan con corte estadístico cuya aplicación no ha sido utilizada aún en los sistemas de extracción de textos, sino a sistemas de Vigilancia Tecnológica. El tema del Procesamiento del Lenguaje Natural se ha venido desarrollando de forma aislada en especialidades como: Ciencia de la Información y Ciencia de la Computación. Es innegable la existencia de

herramientas y procedimientos desarrollados por diversos centros como el CENATAV, que integran el núcleo de investigación nacional en esta temática. Si de aplicaciones se habla, sólo CorpusMiner se perfila como un producto, está desarrollado por la UCLV, desarrollado por Arco (2007).

Los autores más prolíferos en el tema resumen automático, son Aurora Pons, de la Universidad de Oriente, Alfredo Simón del Instituto Superior Politécnico" José Antonio Echeverría" y Leticia Arco García de la Universidad Central "Marta Abreu" de las Villas.

En la Ciencia de la Información el tema es aún menos desarrollado. La primera investigación en el tema de resumen automático data de la década de los años 80 cuando se desarrolló la tesis doctoral de la Dra. María Teresa Cabada (Cabada, n.d.), la cual fue el primer acercamiento a este tema desde la BCI (Bibliotecología y Ciencia de la Información). El tema se retomó luego en pleno siglo XXI con los trabajos de Hernández (2007), Castillo (Castillo, 2008) y Leiva (Leiva, 2008). La investigación de Hernández (2007) devela los retos y los paradigmas en los que se inserta la Representación de la Información, la de Castillo (2008) aborda las tendencias del resumen en el terreno de la BCI (Bibliotecología y Ciencia de la Información) y es un recorrido por el desarrollo del resumen en las organizaciones de información. El trabajo de Leiva (2008) es una metodología para extraer texto de forma semiautomática. Estas investigaciones empiezan a mostrar el desarrollo del tema en la BCI (Bibliotecología y Ciencia de la Información). Otras ciencias como la lingüística parecen estar esperando por entrar en esta área de conocimiento y no muestran avances en este terreno. Unido a esta situación en el plano nacional se aprecia la escasa visibilidad de las pocas investigaciones que se realizan y la poca asociación entre las entidades, es decir poca cooperación institucional. En opinión del autor las dificultades que tiene hoy esta área de investigación en Cuba se da por las situaciones que se plantean a continuación:

- Escaso nivel de integración entre los investigadores que asumen esta temática.

- La producción de software sólo muestra herramientas estadísticas de poca aplicación social.
- Desconocimiento de las potencialidades de estas herramientas para el trabajo de las Organizaciones.
- Escasa visibilidad de las investigaciones en este terreno.
- La formación académica no se detiene en aspectos de la representación automática y asumen otras líneas más tradicionales y menos complejas.
- Bajo nivel de integración entre la producción de software y las necesidades de las organizaciones

### *3.12.- Consideraciones sobre los métodos de Resumen Automático*

Pese a que tenemos gran cantidad de herramientas y software que realizan el Procesamiento del lenguaje Natural, aún es imposible construir de forma automática un texto coherente, sin embargo hay elementos alentadores que reflejan que el camino hacia la cohesión en los resúmenes está al aparecer. Evidencia de ello son los siguientes aspectos:

- Logro una mejor cohesión a través de los elementos conectivos, es decir explotando la hipertextualidad.
- Desarrollo y explotación de las estructuras textuales.
- Explicitación de los elementos estructurales de los documentos fuentes para que coexistan en el resumen elementos extractados y partes relevantes de la fuente original.
- Presencia del documento resumen en formato científico, es decir explicitación de los segmentos textuales acordes a dominios científicos.
- Posibilidad de elaborar resúmenes ajustados a diversos dominios del conocimiento.

Se ha logrado un notorio desarrollo de las técnicas cibernéticas de tratamiento textos. Si bien las primeras técnicas de tratamiento textual estaban ancladas en principios estadísticos, hoy los procedimientos buscan el análisis de las estructura del texto tratando de examinar los contextos. Los criterios eminentemente léxicos se han sustituido por criterios léxicos, semánticos y pragmáticos. Moreiro (Moreriro, 2004) precisa que es necesario desarrollar corpus terminológicos si se quiere desarrollar algunas actividades como:

- Ordenar Conocimientos Científicos.
- Formular informaciones específicas.
- Transferir el conocimiento en la enseñanza.
- Recuperar e Indicar información.
- Sintetizar información.

También se hace necesario el desarrollo de otro elemento esencial en las técnicas sumistas: los analizadores, instrumentos imprescindibles para el desarrollo de cualquier modelo lingüístico automatizado, los cuales se componen de diversos factores estructurales como:

- Morfología
- Lexicografía
- Modelos algorítmicos de sintaxis
- Redes semánticas
- Interpretación o inferencias: Los marcos, las anáforas, etc., todo lo que pueda llegar a la eliminación de las ambigüedades. Desde el factor cognitivo: la función personal, lo subjetivo, el nivel comunicativo, la situación comunicacional y el contexto. También desde la pragmática se debe analizar la retórica y lo convencional de cada idioma.

No obstante se presentan problemas que hoy tienen que resolverse desde un paradigma multidisciplinar y transdisciplinar, entre ellos se encuentra:

- La operacionalidad de los sistemas de extracción de textos sólo en campos restringidos.
- La obtención de extractos inconexos debido a la discordancia que se da entre los elementos cohesivos del texto y sus referentes.
- La ausencia de corpus detallados y amplios para situar de forma contextual el saber de cada disciplina lo que impide contextualizar los asuntos específicos de cada original.
- La definición de mecanismos de análisis de dominio con capacidad de explicitar información oculta, es decir uso de herramientas de estudio de necesidades capaces de revelar los elementos que el usuario no comunica.

## *3.13.- Aportes de la Ciencia de la Información para desarrollo de los Sistemas de Resúmenes*

La Ciencia de la Información como disciplina integradora de diversos procedimientos se imbrica de diversos saberes. El autor de esta investigación reflexiona sobre la base de algunos estudios que en el campo teórico de las técnicas de representación de la información (Descritos desde la ciencia de la Computación, Ciencias de la Información y la Lingüística) se han venido construyendo y que están ejerciendo influencias en el terreno del resumen automático.

Hay elementos que ya tienen una mirada diferente desde la BCI (Bibliotecología y la Ciencia de la Información) y que demandan una proposición práctica en el terreno del resumen automático, campo que necesariamente está demandando el desarrollo de nuevas técnicas de algoritmación, modelación de dominios, ergonomía, extracción y representación de conocimiento. Para este autor la Ciencia de la Información en el entorno de los resúmenes puede aportar disímiles visiones teórico prácticas y en función de los referentes en que se centra esta investigación expone las reflexiones siguientes:

- El análisis de dominio como herramienta para desarrollar el estudio de las comunidades a las que se destinan los productos, lo cual representa un modelo de estudio bien detallado que sólo puede describirse desde la Ciencias de la Información. La noción de dominio en el ámbito de la Ingeniería de Software y la Inteligencia Artificial están limitadas como herramientas para estudiar grupos o comunidades, según Ramírez (Ramírez, 2007). En este mismo campo aparece la introducción del Análisis de discurso como instrumento de modelación de una comunidad epistémica, hecho que si bien está descrito teóricamente, ya empieza a tomar cuerpo en diversas investigaciones desarrolladas desde la Ciencia de la Información y la Ciencia de la Computación, desarrollado en los trabajos de Stacy (Stacy, 2007) y Zadua (2006).
- La visión sistémica en los sistemas de resumen debe sustentarse en el holismo. Los principales sistema de minería de texto funcionan acorde al Ciclo de vida de la Información, sus etapas esenciales son: procesamiento, almacenamiento, difusión, búsqueda y recuperación de información con un fuerte componente sistémico y sinérgico (Peralta, et. al., 2006). El enfoque de estos autores se centra en que los retos de los sistemas de representación de textos tienen una convergencia operacional con el Ciclo de Vida de la Información.
- Dentro de la Lingüística Documental están apareciendo los vocabularios altamente detallados y populares (folksonomías) a partir de la posiciones prácticas y teóricas de Taube, que han sido desarrolladas a través de relaciones de grafos, pero que en esencia son las mismas aportaciones realizadas por la Ciencia de la Información, redimensionadas desde otra disciplina, esto le permite a la lingüística computacional la construcción de corpus detallados o bases de conocimiento (Arco, 2007). El análisis de la dimensión de los vocabularios controlados es una posición que la Ciencia de la Computación ha venido asumiendo como herramienta de comunicación propia, sin embargo la reflexión de esta autor se centra en

una posición donde los preceptos de la matemática se mezclan con los de la Ciencia de la Información.

- El tratamiento retórico de los textos a partir de estrategias de análisis netamente cognitivista que centran su base en el análisis de dominio Zaldua (2006). Esta ha sido una proyección que si bien se describió en los años 90 con los trabajos de Man y Thompson (Mann and Thompson, 1986.), no es hasta este siglo en que comienza a dar frutos. Zaldua (2006), expone los derroteros que debe asumir el análisis de discurso para la Representación de la Información.
- Uso de los procedimientos de categorización como herramientas de representación de contenidos ante la inexistencia de metadatos para representar los resúmenes en los sistemas, lo que demuestra que las necesidades en este tipo de investigación demandan de la unión de dos procedimientos; extracción y abstracción.
- Nuevas formas de tratamiento de la ambigüedad para lograr la calidad en los extractos: sistema de resolución de anáfora pronominal.
- Organización de conocimiento a través de sistemas clásicos utilizados en la Ciencia de la Información, Ej.: taxonomías.
- Aprovechamiento de las folksonomías para mejorar las vías de acceso a la información.
- Desnormalización de los procesos representacionales para hacerlos más abiertos de acuerdo al paradigma cognitivista de la Ciencia de la Información.
- La Formulación de modelos que propicien la confección de agentes, algo que la Ciencia de la Computación ha logrado desarrollando herramientas de investigación de la Ingeniería en Control Automático, Ej.: Modelo (Endres-Niggemeyer, 1995) formulado a partir de los criterios de María Pinto Molina. Esto se fundamenta en que sin dudas cada forma de

representación se convierte en la imagen de pensamiento de la comunidad epistémica a la que pertenece.

- Vinculación con la iniciativa de Archivos abiertos para que los documentos sean almacenados y reutilizados en los procesos de búsqueda, esto posibilita obtener repositorios de resúmenes.
- Utilizar los procedimientos matemáticos de agrupamiento para describir nuevas formas de contenido.

Esto demuestra la necesidad de nuevos métodos y procedimientos sumistas desde una perspectiva diferente donde la lingüística, la Ciencia de la Computación y la Ciencia de la Información se integren, dándole a los contextos, los dominios y las bases de conocimiento léxico un papel preponderante.


# Referencias Bibliográficas

### *3.13.- Referencias Bibliográficas*

AGGARWAL, C. & YU, P. 2005. Online analysis of community evolution in data streams. *In Proceeding of SIAM* SIAM International Data Mining Conference.

AGGARWAL, P. & MUSTAFA, N. 2004 K-means projective clustering. *PODS.* Paris, France: ACM Press.

ALONSO, L. & FUENTES, M. 2002. Collaborating discourse for Text Summarisation *Proceedings of the Seventh ESSLLI Student Session. .* Trento.

ALLEN, M. 2005. *Estructuras de Datos y Algoritmos*, Addison-Wesley Iberoamericana.

AMARO, L. 1977. *La redacción y el resumen*.

ANDERSEN, J. 2006. Knowledge organization: a sociohistorical analysys and critic. *Consulting Library Quarterly,* 76**,** 300-332.

ANKERST, M., BREUNIG, M., KRIEGEL, H. & SNADER, J. 1999. OPTICS: Ordering Points To Identify the Clustering Structure. *In Proceding of In International Conference on Management of Data Mining.* Philadelphia, PA, USA: ACM Press.

ARCO, L. 2007. *Corpus miner: herramienta para el etiquetado de grupos y la obtención de extractos.* MSc., Universidad Central de las Villas.

ARCO, L. 2008. *Agrupamiento basado en intermediación diferencial.* Tesis Doctoral, Universidad Central"Marta Abreu" de las Villas.

ARETOULAKI, M. 1996. *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis Of Texts For Their Automatic Summarization. .* Tesis doctoral University of Manchester.

ARETOULAKI, M. 1997. COSY-MATS: An Intelligent and Scalable Summarisation Shell. *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madrid.

ASLAM, J., PELEKHOV, K. & RUS, D. 1998. Static and dynamic information organization withstar clusters. *In Proceeding of Conference of Information Knowledge Management.* Baltimore.

AUSTIN, J. L. 1962. *How to do things with words. ,* Oxford: , Clarendon Press.

BACKER, F. & HUBERT, L. 1976. A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. *Journal of the American Statistical Association,* 71**,** 870-878.

BALDWIN, B., DONAWAY, R., HOVY, E., LIDDY, E., MANI, I., MARCU, D., MCKEOWN, K.MITTAL, WHITE, M. V., MOENS, M., RADEV, D., SPARCK-JONES, K., SUNDHEIM, B., TEUFEL, S. & WEISCHEDEL, R. 2000. An Evaluation Road Map for Summarization Research. *The Summarization Roadmap*.

BARZILAY, R. 1988. Summarization evaluation methods: Experiments and analysis. *Proceeding of AAAI Intelligent Text Summarization Workshop.*

BARZILAY, R. & ELHADAD, M. 1999. *Using Lexical Chains for tex Summarization* [Online]. Negev. Available: http://acl.ldc.upenn.edu [Accessed 2008].

BELLO, R., ARCO, L. & ARTILES, M. 2006. New clustering validity measures based on roughset theory. *In:* FALCÓN, R. & BELLO, R. (eds.) *In Proceeding of International Symposium on Fuzzy and Rough Sets (ISFUROS'06).* Santa Clara, Cuba: Universidad Central"Marta Abreu" de las Villas, Facultad de Matemática Física y Computación.

BERGER, A. & MITTAL, V. 2000. A system for summarizing Web Pages. *Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval. .* Atenas.

BERGLER, S. 2004. *Multi-ERSS 2004, The CalC Laboratory* [Online]. Concordia: Departament of Computer Science. Concordia University.: Disponible en:http://www.nlpir.nist.gov [Accessed 26.feb. 2009].

BERRY, M. 2004. *Survey of Text mining: Clustering, Classification, and Retrieval,* New York, USA, Springer Verlag.

BOGURAEV, B. & KENNEDY, C. 1997. Salience-based content characterization of text documents. *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madrid. .

BOLELLI, L., ERTEKIN, S., ZHOU, D. & GILES, C. L. 2007. A clustering method for web data with multi-type interrelated components. *In Proceedings of 16th international conference on World Wide Web.* ACM Press.

BOLEY, D. 1988. Principal Direction Divisive Partitioning. *Data Mining and Knowledge Discovery,* 2**,** 325-344.

BORDES, A., ERTEKIN, S., WESTON, J. & BOTTOU, L. 2005. Fast kernel classifiers with onlineand active learning. *Journal of Machine Learning Research***,** 1579-1619.

BRNADOW, R. & MITZE, Y. 2004. Automatic condensation of electronic publication by sentence selection. *Information Processign Management,* 43**,** 5,675-685.

CABADA, M. n.d. *[Automatización del servicio de resúmenes].* Tesis Doctoral, Universidad de la Habana.

CASTILLO, M. D. 2008. *Tendencias del resumen automático.* Universidad de la Habana.

CORTES, C., PREGIBON, D. & VOLINSKY, C. 2001. Communities of interest. *In Proceedings of 4th International Conference on Advances in Intelligent Data Analysis.*

CHENG, D., KANNAN, R., VEMPALA, S. & WANG, G. 2006. A divide-and-merge methodology for clustering. ACM Trans. *Database Syst,* 31**,** 1499-1406.

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DEODHAR, M. & GHOSH, J. 2007. A framework for simultaneous co-clustering and learning from complex data. *In Proceedings of 13th ACM SIGKDD*

*International Conference on Knowledge Discovery and Data Mining.* San Jose, California, USA: ACM Press.

DOURISBOURE, Y., GERACI, F. & PELLEGRINI, M. 2007. Extraction and classification of densecommunities in the web. *In Proceedings of 16th international conference on World Wide Web.* . Banff, Alberta, Canada: ACM, Press.

DUNNING, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics,* 19.

EDMUNDSON, H. 1969. New Methods in automatic extracting. *Journal of the Asociation of Machinery,* 16**,** 264-285.

EDMUNDSON, H. 2006. *Methodology of abstracting Science,* Austin, Texas.

ENDRES-NIGGEMEYER, B., MAIRE, E. Y SIGEL, A. 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing & Management,* 31**,** 631-674.

EPTER, S. & KRISHNAMOORTHY, M. 1999. A multiple-resolution method for edge-centric dataclustering. *In Proceedings CIKM 1999 International Conference on Information andKnowledge Management.* Kansas City, Misoury: ACM Press.

FALKOWSKI, T., BARTELHEIMER, J. & SPILIOPOULOU, M. 2006. Community Dynamics Mining. *In Proceedings of 14th European Conference on Information Systems.*

FERNÁNDEZ, S., SANJUAN, E. & TORRES-MORENO, J. M. 2007. Énergie textuelle de mémoires associatives. *Actes de la conférence Traitement automatique des Langues Naturelles.* Tolouse.

FERRER, R. & SOLÉ, R. 2001. The small wolrd of human language. *Proc. R. Soc. Lond. B,* 268**,** 2261-2265.

FORTUNATO, S., FREEMAN, L. & MENCZER, F. 2006. Scale-free network growth by ranking. *Physical Review Letters,* 96**,** 218701.

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.

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. .

FUKUMOTO, J. 2003. Text summarization based on itemized sentences and similar parts detection between documents. *In Proceedings of Third NTCIR Workshop.*

GAIZAUSLAS – WILKS, J. 1988. Sistemas de trabajo con sumarización. Santa Clara, Cuba.

GAO, B., LIU, T.-Y., ZHENG, X., CHENG, Q.-S. & MA, W.-Y. 2005. Consistent bipartite graphco-partitioning for star-structured high-order heterogeneous data co-clustering. *In Proceedings of 11thACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* Chicago, USA: ACM Press.

GARVEY, S. & GRIFFIT, A. 2008. *[Algunas herramientas de los estudios de usuarios]* [Online]. Santa Clara: Unversidad Central de las Villas. Available: http://intranet.cdict.uclv.edu.cu [Accessed abril.25 2009].

GIBSON, D., KLEINBERG, J. & RAGHAVAN, P. 1998. Clustering categoricl data: an approach based on dynamical systems. *In Proceedings of 24th International Conference on Very Large Data Bases.* New York,USA: Morgan Kaufmann.

GIL-GARCÍA, R., BADÍA-CONTELLES, J. & PONS-PORRATA, A. 2003. Extended Star clusteringalgorithm. *In Proceedings of CIARP.*

GOLDSTEIN, J. 1999. Summarizing Text Document: senetence Selection and Evaluation Metrics. *In Proceeding of SIGIR'99.*

GOLDSTEIN, J. 2000. *Creating and Evaluating Multi-document Sentence*.

GOTLIEB, G. & KUMAR, S. 1968. Semantic clustering of index terms. *Journal of the ACM (JACM),* 15.

GOWER, J. & ROSS, G. 1969. Minimum spanning trees and single-linkage cluster analysis. *Applied Statistics,* 18**,** 54-64.

HALLIDAY, M. & HASAN, R. 1976. *Cohesion in English,* Essex, Longman.

HAN, J. & KAMBER, M. 2001. Data mining: concepts and techniques.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

HJØRLAND, B. 2004. Domain analysis in information science. *Encyclopedia of Library and Information Science.* New York: Marcel Dekker.

HOCHBAUM, D. & SHMOYS, D. 1985. A best possible heuristic for the k-center problem. *Mathematics of Operations Research,* 10**,** 180-184.

HU, X. & WU, D. 2007. Data Mining and Predictive Modeling of Biomolecular Networkfrom Biomedical Literature Databases EEE/ACM Trans. *Comput. Biol.Bioinformatics,* 4**,** 251-253.

INGWERSEN, P. 1992. *Information retrieval interaction,* London, Taylor Graham.

JACKSON, P. 2001. *Multidocument text retrieval,* New York, Amblin.

JACKSON, P. & MOULINIER, I. (eds.) 2002. *Natural Language Processing for Online Applications.* John Benjamins Publishing Company.

JAIN, A. & DUBES, R. 1988. *Algorithms for clustering data,* Englewood Cliffs, NJ, Prentice Hall College Div.

JAMES, A. & GUPTA, R. 2001. Information Retrieval for a summarizing. Center for Intelligent Information Retrieval Temporal Summaries of News Topics

JONYER, I., COOK, D. & HOLDER, L. 2002. Graph-based hierarchical conceptual clustering. *Journal of Machine Learning Research,* 2**,** 19-43.

KALISKY, T., BRAUNSTEIN, L. A., SREENIVASAN, S., BULDYREV, S. V., HAVLIN, S. & STANLEY, H. E. 2006. Scale-free networks emerging from weighted random graphs. *Physical Review E,* 73**,** 025103.

KAUFMAN, L. & ROUSSEEUW, P. 1990. *Finding groups in data: an introduction to cluster analysis*, John Wiley.

KRIEGEL, H. & PFEIFLE, M. 2005. Density-based clustering of uncertain data. *In Proceeding of 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* Chicago, Illionais, USA: ACM Press.

KUPIEC, J. 2003. A trainable Document Summarizer. *In Proceeding of 18 th Annual International ACM SIGUIR Conference og rsearch Development in Information Retrieval.*

KUPIEC, J., PEDERSEN, J. O. & CHEN, F. 1995. A trainable document summarizer. *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95).* Seatle.

LEIVA, A. 2008. *Metodología para la extracción y desambiguación de textos científicos.* Tesis de Maestría, Universidad de la Habana.

LEIVA, A., SENSO, J., DOMÍNGUEZ, S. & HÍPOLA, P. 2009. An Automat for the semantic processing of structured information. *In ISDA 9na International Conference of Desing of Software and Aplicactión.* Italia, Pissa: IEEE.

LIN, C. & HOVY, E. 1998. Automatic Evaluation of summaries using n-gram co-ocurrence Statistic. *In Proceeding of HLTNAACL.* EE.UU.

LIU, Y., CAI, J., YIN, J. & HUANG, Z. 2006. An efficient clustering algorithm for small textdocuments. *In Proceeding of Seventh International Conference on Web-Age Information Management (WAIM 2006). .* IEEE Commnunications Society.

LUNH, H. 1958. The Automatic creation of Literature abstracts. *Journal of Research of Development***,** 159 – 165.

MANI, I. & BLOERDON, E. 1998a. Machine learning of Generic and User-focused. *In Proceeding of the Fifteenth National Conference on Artificial Inteligence.* ACM Press.

MANI, I. & BLOERDON, E. 1998b. Multi-document Summarization by Graph Search and Matching. *In Proceedings of AAAI.* ACM Press.

MANN, W. & THOMPSON, S. 1990. Rhetorical structure theory: a theory of text organization.

MANN, W. C. & THOMPSON, S. A. 1986. Assertions from Discourse Structure. *In:* CONFERENCE, H. L. T. (ed.) *Workshop on Strategies Computing Natural Language.* Marina del Rey, California.

MANNING, C. D. & SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing,* Cambridge-London, The MIT Press, Foundations of Statistical Natural Language Processing,

MARCU, D. 1997. The Rhetorical Parsing of Natural Language Texts. *In Proceeding of 35th Annual Meeting of the Association for Computational Linguistics, (ACL'97/EACL'97).* Madrid: ACM

MARCU, D. 1999. Discourse trees are good indicators of importance in text. *In:* MANI, I., MAYBURY, M. (ed.) *Advances in Automatic Text Summarization.* MIT Press.

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

MARINELLI, D. 2002. Sistemas para el desarrollo de algoritmos.

MÁRQUEZ, A. 2002. Aprendizaje automático y procesamiento del lenguaje natural. *In:* MARTÍ, M. A. & BOIX, A. L. (eds.) *Tratamiento del lenguaje natural.* Barcelona: Edicions de la Universitat de Barcelona.

MATEO, P., GONZÁLEZ, J. C., VILLENA, J. & MARTÍNEZ, J. L. 2003. Un sistema para resumen automático de textos en castellano. *Procesamiento del Lenguaje Natural,* 31**,** 29-36.

MATHIS, B., RUSH, J. & YOUNG, C. 1973. Improvement of automatic abstracts by the use of structural analysis. *JASIS,* 24**,** 101-109.

MATLCLAF, W. 2006. *A Biblography Research in text Summarization* [Online]. Available: Disponible en: http://www.si.umich.edu/~radev/summarization/large-bib.doc [Accessed 23.marzo.2006 2006].

MCKEOWN, K. & RADEV, D. 1995. Generating summaries of multiple news articles. *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95).* Seattle.

MEDINA, J., HECHEVARRÍA, J. & GONZALEZ, B. 2007. Experimentación de algoritmos. *In:* MANSO, R. (ed.) *Sitio 2007.* Santa Clara: CDICT.

MEDINA, J. & PÉREZ, A. 2007. ACONS: a new algorithm for clustering. *CIARP.*

MIZARRO, S. & TASO, C. 2004. Ephemeral and persistent personalization in adaptive information acces to Scholarly publications on the Web. *Segunda Conferencia internacional de Hipermedia Adaptativa.* Málaga.

MONTALVO, C. 2006. *Evaluacion de la Seleccion, Traduccion y Pesado de los Rasgos para la Mejora del Clustering Multilingüe* [Online]. Available: Disponible en: http://www.CMPI.2006.pdf [Accessed 25.mayo 2008].

MONTEJO-RÁEZ, A. 2005. Algoritmos de alta densidad.

MOREIRO, J. 2004. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural,* Madrid, Ediciones Trea.

MOREIRO, J. 2006. *El resumen científico en el contexto de la teoría de la documentación. Texto y descripción sustancial* [Online]. Madrid. Available: Disponible en:http://www.ucm.es [Accessed 26.octubre 2006].

NOMOTO, J. & MATSUMOTO, S. 2001. A New Approach to Unsupervised Text Summarization. *Proceedigs of SIGIR.*

NÚÑEZ, I. 2005. *AMIGA.* Tesis Doctoral, Universidad de la Habana.

ONO, K., SUMITA, K. & MIIKE, S. 1994. Abstract generation based on rhetorical structure extraction. *Proceedings of the International Conference on Computational Linguistics.* Kyoto.

ORASAN, C. & BOROVETS. 2007. Pronominal anaphora resolution for text summarisation. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007). .*

ORDOÑEZ, C. & OMIECINSKI, E. 2002. FREM: fast and robust EM clustering for large datasets. *CIKM '02 (eleventh international conference on Information).*

ORLANDIC, R., LAI, Y. & YEE, W. 2005. Clustering high-dimensional data using an efficient and effective data space reduction. *In 14th ACM*

*International Conference on Information and Knowledge Management.* Bremen, Germany: ACM Press.

PAICE, D. 1988. The automatic generation of abstracts of technical papers. Lancester: Lancester University Press.

PÉREZ, A. & PAGOLA, J. 2007. Documentación de Gstar y ACONS. La Habana: CENATAV.

PÉREZ, J. & MEDINA, J. 2007. A Generalized Star Algorithm for Clustering. *In:* MANSO, R. (ed.) *SITIO 2007.* Santa Clara, Cuba: Editorial Samuel Feijoó.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez

PONS, A. 2006. Una panorámica de la construcción de extractos de un texto. *Revista Cubana de Ciencias Informáticas,* 2**,** 55-67.

QUIAN, Y., ZHANG, G. & ZHANG, K. 2004. FA\ÇADE: a fast and effective approach to thediscovery of dense clusters in noisy spatial data. *SIGMOD '04:2004 ACM SIGMOD International Conference on Management of Data.* Paris, France: ACM Press.

RADEV, D. 2001. Experiment in single and multidocument summarization using MEAD. *1st Understanding Conference.*

RADICCHI, F., CASTELLANO, C., CECCONI, F., LORETO, V. & PARISI, D. 2004. Defining and identifying communities *Networks. PNAS.* USA: National Academy of Science.

RAMÍREZ, Z. 2007. *El Análisis del dominio en la organización y representación del conocimiento.* Diploma de Estudios Avanzados, Universidad de Granada.

RAMÍREZ, Z. & MONTES DE OCA, A. 2004. *Redes de información,* La Habana, Félix Varela.

RAU, L. 1987. Knowledge Organization and Acces in a Conceptual Information System. *Information Processing and Management,* 23**,** 419-428.

RUIZ-SHULCLOPER, J., ALBA-CABRERA, E. & SÁNCHEZ-DÍAZ, G. 2000. DGLC a density-basedglobal logical combinatorial clustering algorithm for large mixed incomplete data. *Symposium. IGARSS* IEEE

SALTON, G. 1997. Automatic text structuring and summarization. *Information Processing and Management,* 33**,** 193-208.

SALTON, G. & MCGILLM, M. 1983. *Introduction to modern information retrieval.,* Nueva York:, McGraw-Hill.

SANDERSON, M. 2005. Currated used directed summarization from existin tolls. *International Conference of Information and Knowledge Management.*

SCHÜTZE, H. 1992. Dimensions of Meaning. *Proceedings of Supercomputing '92,.* Los Alamitos, California: IEEE Computer Society Press.

SEARLE, J. 1975. Indirect speech acts. *In:* COLE, P. & MORGAN, J. (eds.) *Syntax and Semantics 3. Speech Acts.* New York: Academic Press.

SHEIKHOLESLAMI, G., CHATTERJEE, S. & ZHANG, A. 2000. WaveCluster: a wavelet-basedclustering approach for spatial data in very large databases. *The VLDB Journal,* 8**,** 289-304.

SILBER, H. G. & MCCOY, K. F. 2000. Efficient Text Summarization Using Lexical Chains. *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000). .* Nueva York.

SPINK, A. 1997. Interaction in information retrieval: selection and effectiveness of search terms. *Journal of the American Society for Information Science***,** 34 -39.

STACY, H. 2007. Task-based evaluation of text summarization using Relevance Prediction. *Information Processing and Management* 43**,** 1482–1499.

STEIN, G., BAGGA, A. & WISE, G. B. 2000. Multi-document summarization: Methodologies and evaluations. *TALN.*

STEVENSON, M. & WILKS, Y. 2000. Large Vocabulary Word Sense Disambiguation. *In:* RAVIN, Y. & LEACOCK , C. (eds.) *Polysemy.*

*Theoretical and Computational Approach,.* Oxford,: Oxford University Press.

STUMPF, M., WIUF, C. & MAY, R. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *PNAS.* National Academy of Sciences USA.

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

TORRE, F. & KANADE, T. 2006. Discriminative cluster analysis. *In Proceeing of ICML '06: 23rd International Conference on Machine Learning.* Pensylvania: ACM Press.

TORRES-MORENO, J. M., VELÁZQUEZ-MORALES, P. & MEUNIER, J. G. 2002. Condensés de textes par des méthodes numériques. *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT).* St. Malo.

VAN DIJK, T. 1978. *La Noticia como discurso: comprensión, estructura y producción de la información,* Barcelona, Paidós.

VAN DIJK, T. 1980. *Estructura y funciones del discurso,* México, D.F., Siglo veintiuno.

VAN DIJK, T. & KINTSCH, W. 1983. *Strategies of discourse comprehension,* Orlando, Fla., Academic Press.

WANG, W., YANG, J. & MUNTZ, R. 1997. STING: a statistical information grid approach tospatial data mining. *In Proceeding of 23rd International Conference on Very Large Data Bases.* Athens, Greece: Morgan Kaufmann.

WASSERMAN, S. & FAUST, K. 1994. *Social network analysis: methods and applications,* Cambridge, Cambridge University Press.

WEI, C., YANG, C. & LIN, C. 2008. A Latent Semantic Indexing-Based Approach to Multilingual Document Clustering. *Decision Support Systems,* 45**,** 606-620.

WILKS, Y., FASS, D., GUO, C. M., MCDONALD, J., PLATE, T. & SLATOR, B. 1993. Machine tractable dictionary tools. *In:* PUSTEJOVSKY, J. (ed.) *Semantics end the Lexicon,.* Dordrecht, Kluwer.

XIONG, H., WU, J. & CHEN, J. 2006. K-means clustering versus validation measures: a data distribution perspective. *12th ACM SIGKDD international conference on Knowledge discovery and data mining.* Philadelphia: ACM Press.

YAROWSKY, D. 1992. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of COLING-92.*

YAROWSKY, D. 1993. One Sense per Collocation. *DARPA Workshop on Human Language Technology,.* Princeton, NJ.

YAROWSKY, D. 1994. Decision Lists For Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. *Proceedings of the 32nd Meeting of the Association for Computational Linguistics.* Las Cruces, NM.

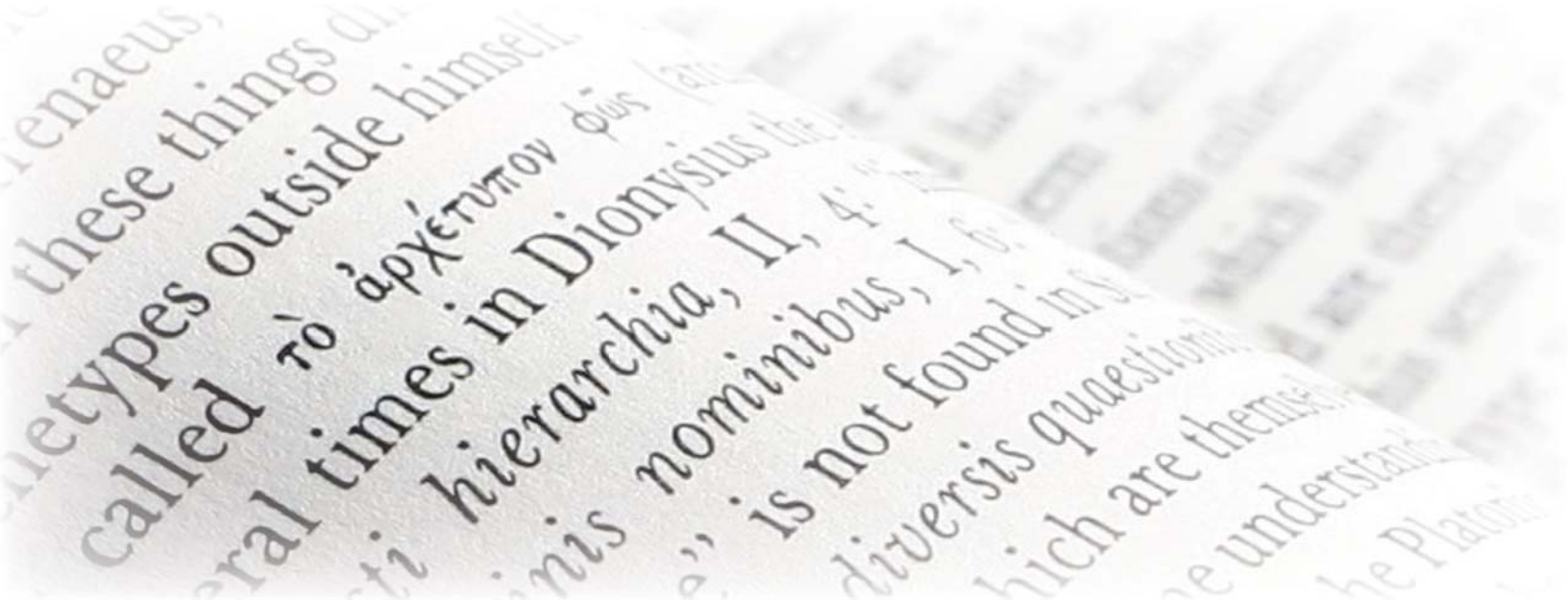
YAROWSKY, D. 2000. Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,,* 34**,** 179-186.

YING, Z. & KARYPIS, G. 2002. Evaluation of Hierarchical Clustering Algorithms for Documen: technical report. Minesotta: University of Minnesota, Department of Computer Science and Engineering.

ZALDUA, A. 2006. El análisis del discurso en la organización y representación de la información-conocimiento: elementos teóricos. *ACIMED* 14.

ZHANG, T., RAMAKRISHNAN, R. & LIVNY, M. 1996. BIRCH: An efficient data clustering method for very large databases. *In Proceeding of International Conference on Management of Data (SIGMOD).* Montreal, QB, Canada: ACM Press.

ZANG, Z., HUANG, Z., ZHANG, X. & 2010. Knowledge Summarization for Scalable Semantic Data Processing. *Journal of Computational Information Systems,* 6**,** 3893-3902.




# Capítulo IV

## Metamodelo para la extracción y desambiguación de textos científicos en el dominio de Ingeniería de Puertos y Costas

# Capítulo 4. Metamodelo para la extracción y desambiguación de textos científicos en el dominio de Ingeniería de Puertos y Costas

## *4.1.- Introducción*

Según Sayão (2001) citado por Hernández (Hernández, 2007) los modelos son esencialmente recursos metodológicos destinados a la adquisición de nuevos conocimientos; son creaciones culturales escogidas para representar la realidad, o algunos de sus aspectos, con el objetivo de tornarlos descriptibles cuantitativa y cualitativamente; son posibilidades de presentar analogías por medio de ciertos formalismos, no solo matemáticos, sino fenomenológicos o conceptuales; forman parte de una dimensión heurística, pues ayudan a develar problemas y a su resolución; y aunque son una estructuración simplificada de la realidad, permiten presentar también las relaciones generalizadas y por tanto, las implicaciones que pueden conducir a nuevas hipótesis y especulaciones.

Las modelizaciones generalmente son una forma calcada de escenarios reales. Según Hernández (Hernández, 2007) son una reducción por relevancia con fines prácticos, y este pragmatismo esencial puede ser descriptivo o normativo. En BCI (Bibliotecología y Ciencia de la Información) se han desarrollado modelos cuyo estatismo está reflejado en el desarrollo de los sistemas metainformativos.

Gracias a la Inteligencia Artificial y de la Ergonomía cognitiva reflejadas en los procesos extractivos se tiene en un lugar prominente la función semántica, sin embargo, según Hernández (Hernández, 2007) aún no se dinamizan los cambios interpretativos en el terreno de la representación de la información, pues los procesos que existen no refieren la forma cambiante de las percepciones y continúan reflejando secciones de la realidad documental pero no su conjunción con el marco ideológico de los signos lingüísticos, porque su relación con la ingeniería del conocimiento todavía es muy instrumental. Las proposiciones metamodélicas desarrolladas por la IA (Inteligencia Artificial) y la Ergonomía Cognitiva no pueden pos sí solas desarrollar modelos extractivos con vistas a la construcción automática de resúmenes. En la BCI

(Bibliotecología y Ciencia de la Información) se encuentran los argumentos metodológicos para la modelación de dominios, hecho que en el campo del resumen ha sido defendido por muchos investigadores como Arco, Ramírez y Hernández (Arco, 2008, Ramírez, 2007, Hernández, 2007) y sin embargo, no se han asumido nuevas formas de hacer para lograr tan importante necesidad. Se habla de comunidades discursivas que deben ser analizadas por análisis de discurso, pero no se refieren qué aspectos del vasto arsenal metodológico del análisis de discurso se utilizará en BCI (Bibliotecología y Ciencia de la Información) y particularmente en el desarrollo del resumen automático.

Las metamodelos para el desarrollo de resúmenes automáticos se erigen como herramientas particulares para el desarrollo de sistemas sumistas, adaptados a diversos dominios. Los principales procedimientos de construcción sumista se han analizado en capítulos anteriores y reflejan el desarrollo de la inteligencia artificial y la minería de texto dentro de este terreno. La Ciencia de la información como disciplina integradora y cosmopolita tiene visiones que sustentan la formulación o el mejoramiento de los métodos establecidos para resumir.

La organización y el análisis de las fuentes de información, la descripción de los metadatos resultantes en los procesos sumistas, el análisis de dominio, la desambiguación y las nuevas formas de acceso al resumen son los elementos que a juicio del autor deben analizarse y reformularse a partir de procesos universalmente descritos, orientados a la lógica de los dominios. En este segmento de la tesis, a partir del análisis documental sobre el tema de la representación textual, el agrupamiento de documentos y el resumen automático de texto, se propone un metamodelo que permite realizar el resumen de textos en documentos con características afines para su ulterior utilización en los sistemas de Gestión de Biblioteca de la Universidad Central "Marta Abreu" de Las Villas.

### *4.2.- Modelo conceptual propuesto que permite el resumen automático de textos*

El procesamiento de documentos textuales debe tomar como punto de partida un modelo conceptual que lo fundamente y cuyo valor metodológico consiste

en integrar un conjunto de conceptos objetivamente interrelacionados, que sirven de base para establecer procedimientos para la obtención de resúmenes extractos de corpus textuales. En conformidad con lo expuesto, en esta investigación se propone un metamodelo que no solo explica el problema científico formulado y también muestra la solución conceptual a éste, sino que también está dirigido a ayudar y orientar a investigadores y desarrolladores en el campo de la minería de textos, la inteligencia artificial y la Ciencia de la Información. Como es característico a todo metamodelo se le definen objetivos, principios, premisas, entradas, salidas, procedimientos y control.

El objetivo del modelo es dotar a los investigadores y desarrolladores en el campo de la Bibliotecología, la Ciencia de la Información, Ciencia de la Computación y Ingeniería en Control Automático, de una herramienta metodológica que posibilite la representación resumida de textos obtenidos de la red de redes a partir de documentos electrónicos que poseen similitud en su estructura y que estén ligados a dominios específicos.

Los principios en que se sustenta el modelo son:

**Consistencia lógica**[21]**.** En consecuencia con el desarrollo de una sucesión de sus pasos en la secuencia planteada y su vinculación en la correspondencia con la lógica de la ejecución de este tipo de estudio.

**Flexibilidad**[22]**.** Por la posibilidad de aplicarse a otras áreas de la minería de textos, la inteligencia artificial, así como en otros procesos que hoy se estudian en la Ciencia de la Información como Vigilancia Tecnológica, Indización automática, etc. con características no necesariamente idénticas a las seleccionadas dentro del universo de estudio y por la capacidad de actualización y reajuste en los diferentes procesos y procedimientos específicos.

**Parsimonia**[23]**.** Es complejo, tiene un alto nivel de sofisticación acorde con los últimos adelantos esbozados generalmente desde la inteligencia artificial, la lingüística y la Ciencia de la Información.

[21] Se refiere a la lógica de los procesos que se describen en la metodología
[22] Capacidad del Metamodelo para adaptarse a diversos entornos.
[23] Se refiere al nivel de complejidad que posee el método propuesto.

**Racionalidad.** Es un proyecto racional acorde con la relación gasto - beneficio que se requiere para su implementación.

**Perspectiva o generalidad[24].** Esto viene dada por la capacidad de este método para conformar sistemas de representación documental, teniendo en cuenta la Minería de Texto, y la Inteligencia Artificial, además esta perspectiva se da en la posibilidad de su extensión como instrumento metodológico para ejecutar estos estudios en otros procesos dentro de la Representación de la Información.

Las premisas fundamentales del modelo se refieren a continuación:

- La colección de documentos en idioma Inglés y español debe estar almacenada en un fichero texto.
- Es necesaria la especificación de bases de conocimiento para la lematización, toquenización, homogeneidad ortográfica, contracciones, abreviaturas y palabras gramaticales y el desarrollo de diversas relaciones de inferencia.

El modelo conceptual reúne todos los elementos considerados relevantes e imprescindibles cuando se pretende por cada funcionalidad brindar varios métodos que la sustenten, teniendo en cuenta uno de los principios en los que se sustenta su flexibilidad. Si se realizan las adecuaciones pertinentes el modelo puede aplicarse al procesamiento de textos en otros idiomas.

La entrada al modelo es una colección de documentos en idioma inglés y español, específicamente artículos de revistas especializadas en Ingeniería. Las salidas principales son la representación resumida del corpus y las referencias URL de los textos. En este metamodelo se toma como referente la propuesta de SIM – SUM que basa su estrategia sumista en un modelo cognitivo – lingüístico. En este modelo Endres-Niggemeyer (Endres-Niggemeyer, 1995, Endres-Niggemeyer, 2005)[25] describe un sistema basado en ingeniería del conocimiento en el cual los agentes inteligentes se sirven de bases de conocimiento diversas, restrictas a dominios específicos. Este agente busca en la base de conocimiento los elementos que se corresponden con el

[24] Es una cualidad que posee el método es su adaptabilidad a otros modelos y sistema
[25] Autores del modelo SIM-SUM

saber que tiene descrito o sea el modelo mental de un resumidor humano (Anexo 37).

Hasta aquí se ha intentado identificar de forma sucinta el marco operativo general, que se pretende utilizar para obtener un resumen abstracto de un corpus textual partiendo de la verificación de la homogeneidad del mismo utilizando métodos semánticos basados en estructura profunda. Se reconoce que pudieran seguirse diferentes variantes con tales propósitos; sin embargo, se ha tratado de captar en el modelo propuesto aquellos elementos esenciales que cualquier investigador o desarrollador en el campo de la Ciencia de la Información debe tomar en consideración al pretender formular Sistemas de Representación de textos de un corpus textual. Como se hizo alusión con anterioridad, al metamodelo se le definen, además, procedimientos que posibilitan su implementación. El modelo está basado en tres modelos esenciales: modelo lingüístico, modelo naturalista y modelo matemático.

### 4.2.1- Modelo Lingüístico

Los sistemas que resuelven el problema que nos proponemos solucionar con este método son los llamados sistema de “corta y pega”, los cuales se elaboran a partir de un léxico detallado que se trata mediante técnicas cibernéticas. Es evidente que para desarrollar estos sistemas se hace necesario contar diversas herramientas y a partir de las mismas realizar múltiples tratamientos textuales entre los que se encuentran los siguientes:

- Desarrollo de sistemas para la resolución de anáfora.
- Construcción de ontologías y diccionarios semánticos a partir de los procederes ideados para Wordned.

Este enfoque metodológico se sale de los criterios de selección textual desarrollados en otras ocasiones, los cuales se suman a los criterios estadísticos establecidos en otros sistemas. Los modelos sumistas anteriores obligan a trabajar con medios lingüísticos que permitan la detección de elementos textuales con carga semántica y al desarrollo de métodos que proporcionen la modelación de las necesidades de la red social a la que están dirigidos. A la frecuencia de aparición de términos en determinado corpus se

añade la aparición de palabras específicas de una comunidad epistémica, lo que evidencia la relación o la aplicación oportuna de estos métodos en dominios restringidos. Los modelos que se insertan en esta vertiente sumista permiten que las oraciones seleccionadas puedan organizarse de acuerdo con una estructura retórica determinada. Según Moreiro (Moreiro, 2004), para lograr una representación coherente y de calidad del texto, estos modelos apelan a los aspectos siguientes:

- Representación Interna del texto, mediante modelos de los elementos que lo componen, es decir por razón del desarrollo y diferenciación de entidades, sintagmas nominales, los nombres propios, etc. y sus relaciones a través de vectores.
- La búsqueda de la similitud: contar las palabras que proceden de la misma forma canóniga o sea las que tienen una misma raíz en el momento de establecer un nexo o marco semántico.
- La proximidad entre las entidades del texto, factor determinante antes de establecer sus posibles relaciones.
- Resolver la cohesión: Las entidades del texto que tienen diversos nexos deben tener algún tipo de asociación semántica.
    a) Concurrencia de palabras: Los términos que poseen igual significado deben desarrollarse o expresarse en un mismo contexto.
    b) Similitud Léxica: Relación entre las palabras de igual significado. Control de la sinonimia.
    c) Co-referencia: Control de los términos sinónimos. Esto permite describir y establecer la coherencia entre las oraciones que se refieren indistintamente a algo que semánticamente es similar. El medio de asociación de las palabras es mediante la referencia anafórica.
- Ajustar las diferencias de las macroestructura textual a las condiciones de las superestructuras de un área determinada.

- Delimitar la superestructura textual para explotar la organización del texto en los procesos de resumen.
- Normalizar la presentación de los abstractos.

  a) Eliminar la reiteración de oraciones.

  b) Lograr la coincidencia del estilo con las cualidades del resumen en Ciencia.

  c) Relación entre la discursividad y la coherencia lógica.

  d) Sustitución de la primera persona del plural por la tercera persona.

  e) Supresión de los elementos subordinados.

  f) Armonización de los elementos y los tiempos verbales.

Estas son las concepciones lingüísticas con las que se apoya este metamodelo.

### 4.2.2. Modelo Naturalista (Enders-Nigguemeyer, 2005)

Sobre este terreno es importante describir el método naturalista de Enders-Nigguemeyer. Teniendo en cuenta que se conoce que el proceso cognitivo está condicionado por factores medioambientales. Con el objetivo de observar su desempeño en condiciones reales y obtener observaciones objetivas, Enders-Niggemeyer (Enders-Niggemeyer, 2005) tiene contacto con sus objetos de estudio (los resumidores y los usuarios) en su vida diaria, visitándolos en casa o en su oficina, como está estipulado en los principios de las investigaciones de campo y las encuestas naturalistas. En la concepción de este método se trabajó con 6 resumidores/indizadores – 4 alemanes y 2 estadounidenses quienes realizaron 9 resúmenes cada uno pensando en voz alta.

Los datos recogidos durante este proceso de pensamiento en voz alta fueron utilizados junto con los documentos originales y notas que se tomaron durante el proceso conjuntamente con los borradores de los resúmenes y las versiones finales.

La interpretación de este proceder siguió modelos cognitivos sobre la comprensión de textos y la adquisición de conocimientos a partir de un texto. En la versión digital de este estudio, una base de datos o base de

conocimientos, que pudiera ser una ontología, reemplaza el saber almacenado en el cerebro humano. Como resultado del experimento el perfil cognitivo de 6 resumidores profesionales resultó ser bien organizado y estructurado. La manera en que un resumidor por sí solo organiza el proceso fue diferente, sin embargo todos procedieron paso a paso, analizando elementos disímiles como: documentos, capítulos, párrafos, oraciones, frases o palabras una a una.

Estos pasos sirvieron de base para la construcción de estrategias intelectuales, teniendo en cuenta los documentos originales, los resúmenes efectuados y el conocimiento disponible en la memoria del resumidor. Se descubrió, además, que los grupos de estrategias intelectuales interactúan con los pasos de trabajo individuales. El resultado fue que los resúmenes hechos por los humanos se integran muy bien con la búsqueda de información.

Los resumidores primero identifican elementos que se ajusten a la tarea o interés que les ocupa en ese momento. Después de concentrarse en un solo elemento, buscan en el documento fragmentos de un tamaño razonable que contengan materiales útiles, usando todo tipo de marcadores como palabras clave o algunas características del diseño de la página, y muchas veces usan datos fuera del texto: como el índice o la tabla de contenido. Después de haber encontrado fragmentos del texto interesantes del tamaño de un párrafo, los resumidores los leen cuidadosamente, posiblemente revisando y buscando lo que no esté muy claro para ellos.

Ellos eligen unidades del tamaño de una oración para sus propias anotaciones. Los resúmenes generalmente se realizan cortando y pegando fragmentos, las construcciones personales no ocurren muy a menudo. El abstracto resultante generalmente se revisa y a veces solo necesitan pequeños detalles para convertirse en un resumen aceptable. (Ver figura 11).

Sin embargo, este tipo de modelo está aplicado a las condiciones restrictas de resumidores de artículos de hematología. El autor cree que los mecanismos de percepción del usuario sobre el texto también son necesarios para desarrollar la summarización. No basta solo con describir o modelar un agente con los mecanismos de confección de resumen. En la opinión del autor, estos mecanismos están dados en determinados segmentos y bajo condiciones de

trabajo subjetivas. Los resumidores se valen de diversas técnicas para confeccionar resúmenes informativos, que dependen de su grado de especialización y dominio del tema que representan. En esta variante sumista se demuestra que los resumidores analizan las secciones del texto a través de diversos modelos estructurados, donde la lectura esquemática del texto tiene una preponderancia esencial en el tratamiento del resumen. Las secciones del texto facilitan el análisis de la retórica o la forma específica de escribir en ese dominio.

Otro aspecto al que esta vertiente sumista le presta especial atención es la interpretación y análisis de la estructura conceptual, la cual permite confeccionar asociaciones de tipo fragmentario, que facilitan instituir correctos análisis de las relaciones de los conceptos dentro del texto. Enders-Nigguemeyer (Enders-Nigguemeyer, 2005) ha sido capaz de reanalizar los modelos de resumen y esbozar una teoría diferente de la que se da en los métodos anteriores, sin embargo los métodos de extracción siguen siendo similares a los de otros sistemas, solo que en este caso hay un componente naturalista o cognitivo que facilita el trabajo.

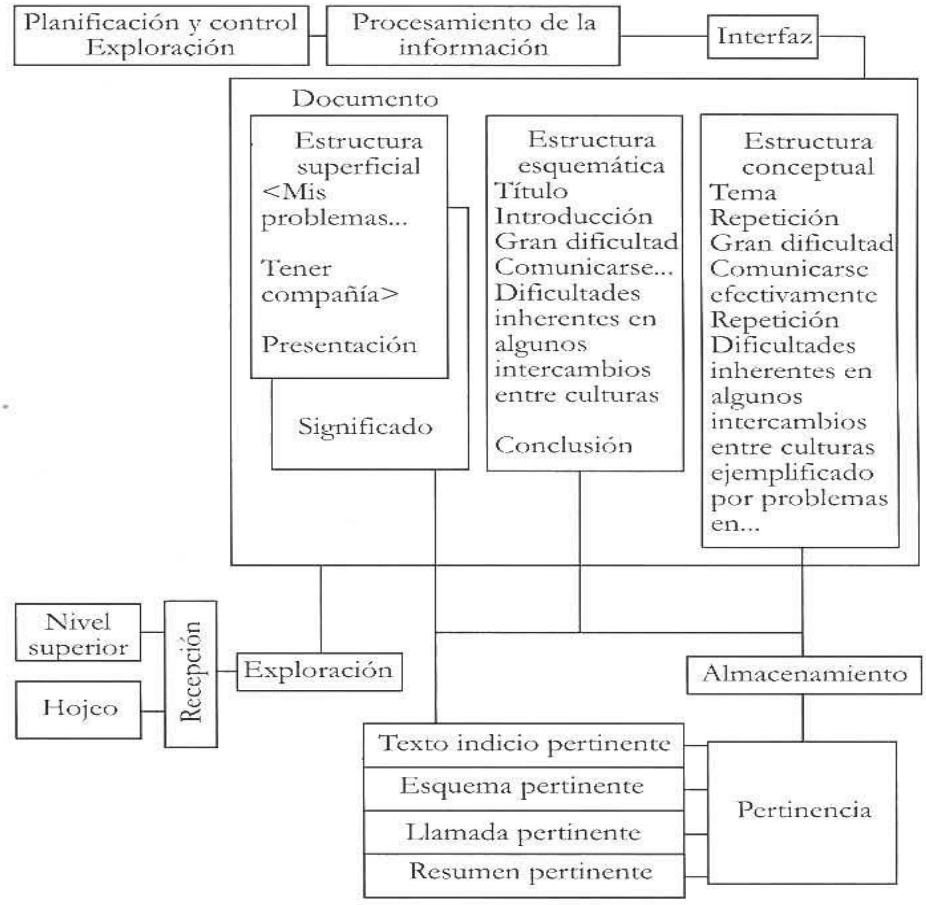



Figura 11: Modelo Conceptual Enders- Nigguemeyer, Pinto (2001)

#### 4.2.2.1.- Concepciones Contextuales: elementos del modelo naturalista (Análisis de dominio)

Estos modelos se centran en formulaciones declaradas inicialmente por Hjørland y Albrechtsen en la década del 90 del pasado siglo. A su vez estos postulados son modelizaciones conceptuales en constante movimiento. En el caso de la investigación que se formula no se estudia un domino temático específicamente, si no una forma específica de mostrar información, o sea se intenta buscar un modelo de organización textual, cuyos referentes toman al artículo científico como centro de análisis.

Según López-Huertas (López-Huertas, 2008) el dominio temático como modelo y referente para el diseño y elaboración de estructuras de conocimiento es un proceder importante para el desarrollo de sistemas de representación de la información individuales, pues permite analizar y estructurar los dominios de

forma independiente. Es por ello que es imposible tratar todos los dominios con los mismos criterios estructurantes. Desde las Ciencias de la Información se considera preponderante el tratamiento de las comunidades de discurso como herramienta diferenciadora, esto hace que se considere el conocimiento del dominio en cuestión como unidad de análisis para la representación y la organización del conocimiento. Esta aseveración viene a mostrar diversas divisiones del análisis de dominio, las cuales tienen diversas formas como:

- ontológicas
- epistemológicas
- sociales

Lo interesante del análisis de dominio y su valor contemporáneo según Hernández (Hernández, 2007) y López Huertas (López-Huertas, 2008) es que manteniendo la mirada en las especialidades científicas como referente, presenta una visión nueva que rompe con la tradición anterior, que pretendía una representación objetiva de la realidad, enfoca los espacios disciplinares en su dimensión social como comunidades de discurso, ampliando y matizando concepciones y fundamentando teóricamente sus posiciones.

El análisis de dominio permite abordar nuevas visiones del individuo como: la organización del conocimiento, sus estructuras, el lenguaje, las formas de comunicación, lo que muestra la actividad de las comunidades de discurso y de su papel en la sociedad. Se reconoce según Hernández (Hernández, 2006) y Ramírez (Ramírez, 2007) que el núcleo del análisis del dominio es el estudio de las actividades y productos de los dominios para tener conocimiento de lo que hay en las estructuras de cada dominio.

Las teorías ontológicas que se desarrollan dentro del análisis de dominio facilitan el conocimiento del mundo y sus objetos, además describen la realidad y su estructura. Para López-Huertas (López-Huertas, 2008) el valor esencial de estas teorías está en su capacidad de reconocer la naturaleza de los fenómenos conocidos, independientemente de los medios utilizados para conocerlos. El principio ontológico constituye una forma intelectual de organización del conocimiento frente a la utilización de principios sociológicos

para la organización del conocimiento, pues en su aplicación subyace una forma detallada del concomimiento.

Es su forma práctica el conocimiento ontológico asume estructura de clasificaciones especializadas, tesauros y ontologías, herramientas útiles en el desarrollo de resúmenes. Todos estos productos se constituyen por conceptos centrales de un dominio determinado y sus relaciones semánticas correspondientes. Estos productos pueden subdividirse de la forma siguiente: según López-Huertas (López-Huertas, 2008):

- Indización y recuperación de la información especializada (Se refiere a la representación de los documentos especializados).
- Estudios empíricos del usuario: Pueden dar información sobre las diferentes necesidades de información de distintas comunidades.
- Estudios bibliométricos: Visualizaciones de los mapas científicos. Evidencia conexiones detalladas entre los documentos.
- Estudios históricos**:** Los métodos históricos deben ser considerados Estudios del documento Su importancia está relacionada con la introducción de sistemas de recuperación de la información a texto completo**.**
- Estudios epistemológicos y críticos: Proporcionan conocimiento sobre los fundamentos teóricos de una disciplina o temática. Dan directrices para la selección, organización y recuperación de la información.
- Estudios terminológicos, semántica de las bases de datos y estudios del discurso.
- Estudios de la estructura y las instituciones en comunicación científica.
- Cognición científica, conocimiento experto e Inteligencia Artificial.
- Aporta modelos mentales de un dominio y métodos para la extracción de conocimiento para generar sistemas expertos.

A las aportaciones teóricas vistas hasta hace un momento se suman los criterios de Ramírez, (Ramírez, 2007) y Hjorland (Hjørland, 2004), en las que se concibe el análisis de dominio como herramienta naturalista capaz de

generar información sobre el usuario, cuyo campo de aplicación tiene una dimensión lingüística, comunicativa y cognitiva. El análisis de dominio permite desarrollar estudios en esferas inherentes al usuario, ellas son el discurso, el contexto y el medio de aplicación, lecturas en la que su sustenta esta tesis (ver figura 12).

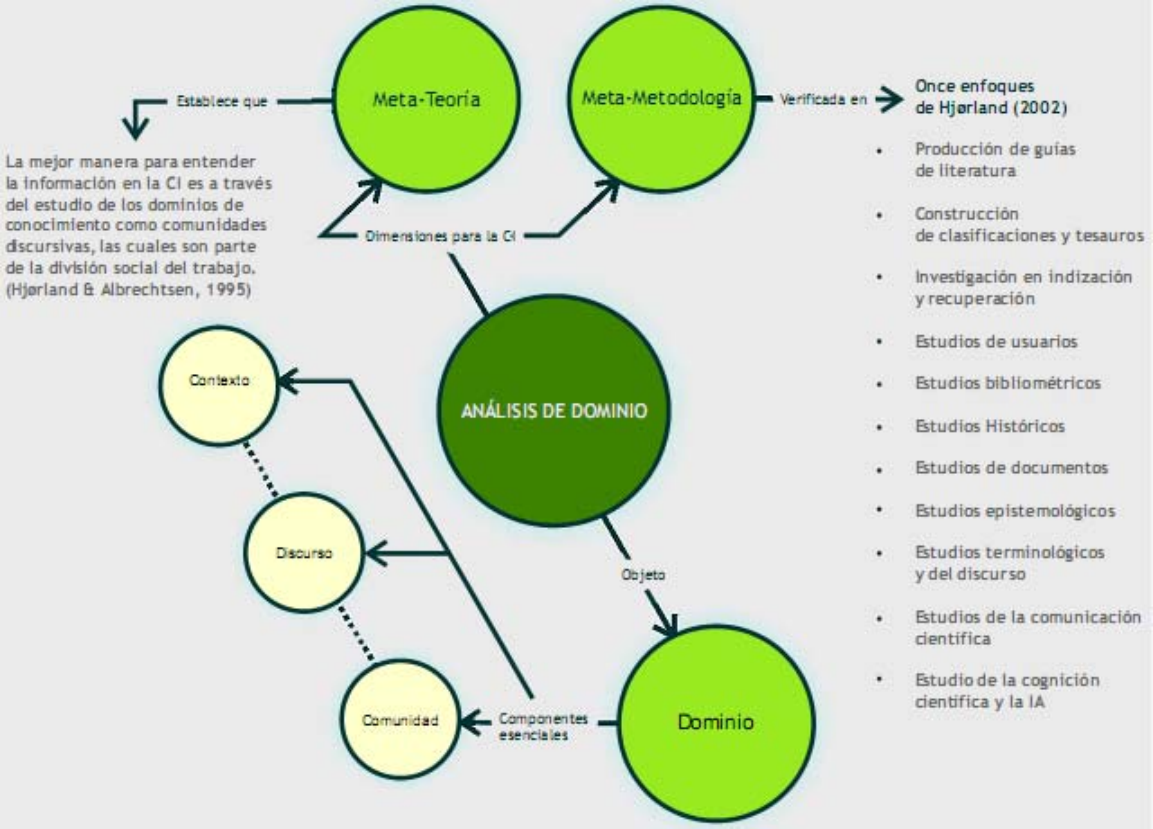



Figura 12. Estructura del Análisis de dominio Ramírez (2007)

### 4.2.3.- Modelo Matemático

Es necesario especificar qué tratamiento matemático se dará a este modelo. Es indiscutible la existencia de múltiples modelos algorítmicos para la construcción de textos. Este acápite de la investigación tiene como objeto identificar los elementos matemáticos con que se comprobará nuestra propuesta y los algoritmos que se aplicarán en los procesos de extracción y clasificación de los documentos estudiados más adelante en la fase experimental. En este estudio se apelará al uso de reglas discursivas desarrolladas a través de agentes siguiendo la línea sociocognitiva de esta investigación, debido a que se trata de explotar al máximo las regularidades encontradas en el discurso objeto de esta

investigación, además las pruebas que se realizarán para constatar la calidad de los experimentos se indicarán de acuerdo a distancias establecidas en el capítulo uno de la investigación. También se aplicarán medidas de evaluación basadas en distancias geométricas.

### *4.3.- Modelo Empírico aplicado a la UCLV*

Atendiendo a las apreciaciones del autor sobre el modelo de Endres-Niggemeyer (Endres-Niggemeyer, 2005) se aplica a los 12 resumidores de las bibliotecas de la Universidad Central "Marta Abreu" de las Villas, observándose en los resultados la coincidencia de los elementos esenciales del modelo, sin embargo en la observación se apreció una tendencia a la interpretación de gráficas y tablas, pues los resumidores son especialistas en las materias que resumieron, por tanto su campo de análisis del resumen integra tablas y formularios unidos a los elementos estructurales descritos en el modelo referido con anterioridad. Los resultados de esta observación y todo el proceso de preparación para la aplicación del instrumento se encuentran la figura 13 y en los Anexos (13 y 14, 1 y 26) (ver figura 13).

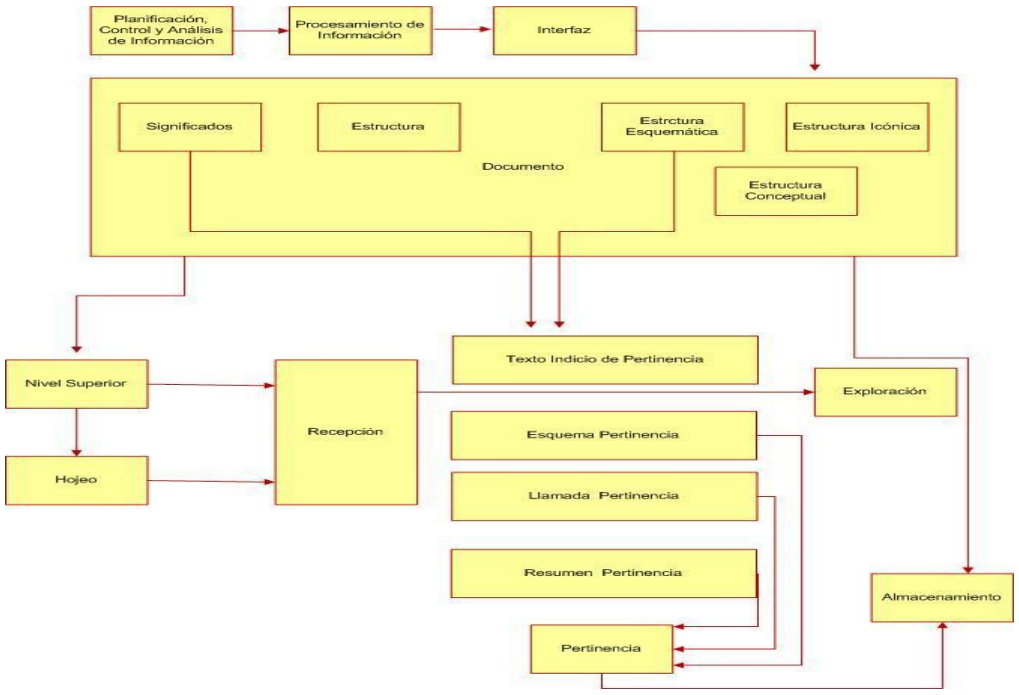



Figura 13. Modelo Cognitivo resultante de la aplicación del Método de Enders-Nigemayer (2003) en la UCLV

### 4.3.1.- Modelo Metodológico Resultante

**Tipo de Resumen:** Monodocumento.

**Forma de Aplicación:** Resumen Basado en Discurso: a partir de las estructuras del discurso del artículo científico se desarrolla este tipo de resumen que funciona con diversas bases de conocimiento y agentes cibernéticos.

### 4.3.2.- El Artículo Científico

**Estructura y contenido**

Como nos interesa resumir artículos científicos en español e inglés se impone analizar la estructura y contenidos habituales de estos textos en diversas publicaciones electrónicas.

**Estructura de los Textos a Resumir**

El espectro del discurso científico está abierto a un gran número de géneros discursivos. Entre estos se encuentran los siguientes: Artículos Originales, Cartas al Director, Editoriales, Revisiones, Conferencias, Proceeding de Eventos, Comunicaciones, etc.), pero en este estudio nos centraremos en el Artículo Original. Las revistas Científicas les solicitan a los autores de artículos que sigan un patrón determinado que debe mantener la estructura siguiente:

- Título
- Resumen en español
- Palabras clave en español
- Resumen en inglés (abstract)
- Palabras clave en inglés (keywords)
- Introducción
- Métodos
- Resultados
- Discusión

- Conclusiones
- Agradecimientos (opcional)
- Bibliografía

**Subtítulos del artículo en las Publicaciones Científicas**

El autor patentiza al igual que D'Cunha (D'Cunha, 2006) que para resumir en forma automática un artículo científico debe buscarse determinado tipo de información de su estructura (Introducción, métodos, Resultados, Casos de estudio, etc.), lo que indiscutiblemente determinará el desarrollo lógico del proceso científico. Esta idea de explotación de la estructura textual en resumen automático también se sigue en el trabajo de Bélanger (Bélanger, 2005).

#### 4.3.2.1.- Unidades léxicas representativas del artículo Científico

Resumir textos en un ámbito especializado conlleva a la detección de las unidades léxicas más importantes que aparecen con cierta frecuencia en dichos textos, ofreciendo indicaciones acerca de cuáles son los contenidos importantes. En el caso que nos ocupa se trataría básicamente de unidades nominales que intuitivamente reflejan conceptos esenciales para un resumen de este tipo (como objetivo, objeto, propósito, intención, resumen, conclusión, resultado) o conjuntos verbales, con frecuencia alta y no auxiliares o demasiado comunes (como asociar, analizar, presentar, evaluar, relacionar, aportar, estudiar, valorar) D'Cunha (D'Cunha, 2006). Mediante la detección de los elementos del artículo científico, el sistema de resumen dispondrá de una primera salida de los conjuntos oracionales (las que incluyen las unidades léxicas que se determinen) que posteriormente se contrastará y refinará mediante otras técnicas lingüísticas que se explican más adelante.

### 4.3.3.- Concepción teórico-metodológica del procedimiento (Semántico Cognitivo) general para obtener el resumen abstracto de corpus textuales

Como parte del modelo conceptual se desarrolla un procedimiento metodológico general que incluye varios procedimientos específicos, estructurados en seis etapas con sus fases correspondientes que en su conjunto resumen el contenido del modelo.

Las etapas del procedimiento general son:

1. Estudio de Necesidades.
2. Análisis Manual del Corpus Textual.
3. Creación de la ontología.
4. Extracción del Texto mediante Agentes Cognitivos.
5. Modelación de sistema de búsqueda y recuperación de información.
6. Representación de la Información.

### 4.3.4.- Etapa 1. Estudio de necesidades

#### 4.3.4.1.- Fase 1. Representación de la comunidad Epistémico (Estudio de Necesidades)

**Confección de las Bases de datos sobre los usuarios del Sistema.**

Se diseña una base de datos con los elementos necesarios del usuario. Esta base de datos registra los documentos que está usando el usuario en su investigación, es el elemento esencial para construir el sistema y para su mantenimiento. Se necesita moldear al usuario de forma lingüística y comunicativa, esto implica que se necesite conocer los términos relacionados con su investigación, así como las Fuentes de información y los recursos que usa la comunidad epistémica a la que se destina el resumen Hjorland (Hjorland, 2004) con el fin de formular un modelo mental de la representación informativa de los usuarios.

A diferencia de los estudios de usuarios desarrollados por Núñez (Núñez, 2005), en este proceso se modela al usuario de acuerdo con su forma de comunicarse e interactuar con las fuentes de información, por tanto el estudio de necesidades para el desarrollo de un sistema de resúmenes no se desarrolla solo a partir de análisis de necesidades de búsqueda de información, pues el entorno digital necesita de elementos cada vez más específicos para el tratamiento de las necesidades, aspecto que va mucho más allá de los estudios de usabilidad Web, pues son estudios esgrimidos sobre elementos que miden la comunicación del usuario con sistemas de navegación visual/lineal-escalable, que en ocasiones sirven para el desarrollo de páginas Web. En este

modelo se busca reconocer las características de los textos para a partir de ellas organizar procesos cognitivos para la organización textual y para el desarrollo de herramientas de organización de contenidos temáticos.

En tal sentido, en este metamodelo se busca representar los grupos o los dominios mediante el análisis de discurso amparado en la Teoría de Van Dijk (Van Dijk, 1984) el cual será combinado con variables clásicas que representen al usuario, no a sus necesidades , estos datos serán obtenidos mediante una encuesta. Estas serán las variables a utilizar:

- Nombre
- Apellidos
- Publicaciones que Utiliza en su Investigación
- Email
- Términos Relacionados con su investigación
- Hora en que necesita el resumen
- Idioma en que prefiere el resumen
- Pasos o acciones que ejecuta al resumir
- Pasos o acciones que ejecuta al leer el resumen.

Técnica de análisis de discurso

**Análisis de la Macroestructura**: Orientada a describir le estructura global de los textos o las fuentes informativas que utiliza el usuario. Este segmento intenta identificar la estructura de los textos a nivel global. Identifica la semántica (el tema de los textos), la sintaxis entre los grupos oracionales, así como la posición pragmática del texto. Estos resultados se introducen en la base de datos que será movida por una ontología.

Para desarrollar este análisis se deben desarrollar las siguientes acciones:

- Identificar las unidades semánticas del texto, estas son las dimensiones hacia donde se enrutan los contenidos.
- Reconocer los elementos que describen elisiones o elaboraciones.

- Marcar los segmentos del texto donde se haga referencia a hechos, procesos y modelos.

La detección de los ejes semánticos se desarrolla a través de la semántica"La organización del discurso no es plana ni lineal; es jerárquica, es decir, las cláusulas forman estructuras de orden superior, párrafos, que a su vez se combinan para formar episodios mayores o secciones del discurso Tomlin (Tomlin, 2003).

El sentido es un aspecto crucial en cualquier descripción del discurso, como algo que los usuarios del lenguaje le asignan en un proceso de comprensión o interpretación, partiendo de que cada texto ofrece una proposición o un conjunto de proposiciones.

Las proposiciones se conectan semánticamente a través de la coherencia, noción responsable de que un discurso tenga sentido, ya sea a nivel micro –entre las proposiciones- o a nivel macro –el discurso en su totalidad-.

Según Van Dijk (Van Dijk, 1984, 1980), en el micronivel las relaciones de coherencia obedecen a una naturaleza funcional (especificación, generalización, ilustración o contraste con respecto a una proposición previa). De igual forma a la hora de buscar coherencia interna o local en un texto podemos hablar de focalización, tópico y referencia.

La descripción del macronivel supera la lingüística y la gramática para asentarse en los tópicos (de qué se está hablando) y los temas, a la vez que se define el sentido de unidad global del discurso, expresado habitualmente en segmentos como los titulares, los resúmenes o las conclusiones.

En relación con el tema global se encuentra la noción de macroestructura formulada por Van Dijk (Van Dijk, 1980). La macroestructura es la estructura semántica global de un discurso y puede expresarse por su título o encabezado o por oraciones de síntesis. Las proposiciones macroestructurales se derivan mediante macroreglas, esto es, mediante la eliminación de aquellas proposiciones que no son pertinentes para la interpretación de otras proposiciones (elisión), mediante la conversión de una serie de proposiciones específicas en una proposición más general (generalización), mediante la

construcción de una proposición a partir de un número de proposiciones del texto (construcción), y a partir del conocimiento activado del mundo Tomlin (Tomlin, 2003).

**Análisis de Microestructura:** Desarrollado para definir las estructuras sintácticas de los elementos gramaticales que se encuentran en las oraciones del texto. Aquí se identifican todas las estructuras gramaticales que deberán ser la primera entrada de términos al sistema de ontologías. Lo esencial en este proceso es lo siguiente:

- Determinar las unidades sintácticas (verbos, sustantivos, pronombres, etc.) estos son la primera entrada de vocablos al lexicón, estos serán formalizados en la ontología.

**Análisis retórico y estilístico:** Para Van Dijk (Van Dijk, 1995), el estilo es el conjunto total de los detalles estructurales variables y característicos del discurso, que son una indicación del contexto social y personal del hablante.

Dado que todo uso del lenguaje se reconoce hoy en día como poseedor de un estilo, sigue siendo cierto que la mayoría de las descripciones de estilo tienen implícita una perspectiva de comparación, de ahí que el estilo se defina, muchas veces por oposición a los demás estilos, es decir, que se defina en la variación. Según Van Dijk (Van Dijk, 1984), cuando analizamos un estilo podemos definir un conjunto de características discursivas típicas de un género, de un hablante, de un grupo humano, de una situación social, de un período literario e incluso de toda una cultura.

El estilo es una dimensión que responde a género, estatus, edad, clase, antecedentes étnicos, las circunstancias, la elección del hablante. Es un perfil específico que funciona en un contexto comunicacional o una dimensión social o grupal comparable con otras situaciones o personas, ya que los hablantes recurren a la variación estilística para expresar significados adicionales o implícitos que resultan pertinentes en la interacción Sanding y Selting (Sandig and Selting, 2003).

Según Ariadna Bolívar “a pesar del tiempo, el texto mantiene sus propósitos comunicativos y sus características generales. Pero se observan diferencias de

estilo y de organización que dependen, fundamentalmente, del tipo de interacción que se da entre el texto y su lector y del sistema de valores que comparte" Bolívar (Bolívar, 1998), razón por la cual el estilo legitima el acto discursivo y viceversa, en un contexto de interacción determinado.

El estilo llega a través de la retórica, que desde el mundo griego clásico fue vista y valorada como la herramienta fundamental para producir discursos. La retórica ha sido asociada también con el arte de persuadir, teniendo en cuenta que el texto o discurso que persuade es aquel que logra finalmente su objetivo gracias a sus propiedades estilísticas.

La retórica se configura como un sistema de reglas y recursos que actúan en distintos niveles en la construcción de un discurso. Tales elementos están estrechamente relacionados entre sí y todos ellos repercuten en los distintos ámbitos discursivos.

Se ocupa, así de sistematizar procedimientos y técnicas de utilización del lenguaje puestos tanto al servicio de una finalidad persuasiva como estética del mismo, añadida a su fin comunicativo.

La retórica estudia tanto las formulaciones como el contexto, a través de ella se pueden explicar los aspectos sociopsicológicos que logran persuadir, basados en sus estructuras, cuyo manejo está relacionado con los objetivos y efectos deseados a la hora de comunicar un mensaje, de ahí que en el nivel cognitivo semántico se pretenda que el receptor entienda lo dicho, lo acepte y realice las acciones pretendidas o ejecute las órdenes implícitas en el texto persuasivo, teniendo en cuenta el carácter retórico de ciertos sucesos y situaciones específicas. "Un texto retórico responde a ciertos temas o problemas propios de una sociedad, o interactúa con ellos, y produce cierta acción o cambio en el mundo." Gill y Whedbee (Gill and Whedbee, 2003).

La persuasión también es el objetivo esencial del discurso científico. En un rejuego con estrategias retóricas propias, tales como la organización jerárquica de la información, la argumentación, las descripciones factuales, la utilización de fuentes y señales de precisión y exactitud como las cifras, la hora, la edad o la comunidad de creencias y conocimiento con el lector refuerzan el valor de los textos científicos y deben ser elementos esenciales para abordar la

metodología asociada a este modelo. Para realizar esta técnica se deben identificar en el texto, unidades de adyacencia como:

- Causa.
- Motivación
- Argumentaciones
- Exactitud en las descripciones.

**Análisis de Contexto:** Según Van Dijk, el contexto desempeña un papel fundamental. Para Van Dijk (Van Dijk, 2004), el estilo es el conjunto total de los detalles estructurales variables y característicos del discurso, que son una indicación del contexto social y personal del hablante.

El enfoque multidisciplinario del análisis discursivo actual busca lograr una visión integradora de sus valores y proposiciones, lo que ha llevado a evaluar cada uno de los discursos en su contexto, como parte constitutiva del entorno local, global, social, cultural, histórico, tanto como la valoración de los procesos de interacción entre los usuarios del lenguaje y las funciones sociales y culturales del discurso, “entender el discurso y su contexto como dos aspectos que se generan y constriñen mutuamente ”Ono (Ono et al., 1994) teniendo en cuenta que el análisis del discurso va mucho más allá de ser una especialidad de la lingüística, debido a que “el contexto suele suministrar indicios sobre algunas expresiones funcionales localmente relevantes que no pueden detectarse cuando solo se consideran los datos lingüísticos. . En la conversación y en los textos hay muchas indicaciones de su pertinencia contextual, lo que obliga a observar y analizar en detalle las estructuras del contexto también como consecuencias posibles del discurso: las situaciones, los participantes y sus papeles comunicativos y sociales, sus metas, el conocimiento social pertinente, las normas y valores, las estructuras institucionales u organizativas, etc.” Van Dijk (Van Dijk, 1980).

En el modelo socio-cognitivo del análisis del discurso defendido por Van Dijk, el contexto explica lo más relevante en la información semántica de un discurso como un todo que distingue valores locales y globales del discurso tales como la situación (tiempo, lugar, circunstancias), los participantes y sus diversos

papeles comunicativos y sociales (hablante, coordinador, amigo), las interacciones, metas, propósitos y el conocimiento; de ahí que se entiendan como estructuras construidas y reconstruidas en el momento por cada participante en un evento comunicativo, lo que se ha dado en llamar una relación dialéctica entre el discurso, sus usuarios y el contexto.

El discurso se localiza en la sociedad como una forma de práctica social o de interacción de un grupo o comunidad epistémica. Los miembros del grupo tienden a interpretar y representar la realidad de acuerdo con sus intereses más importantes, aunque entre el discurso y el contexto se establece una relación de interfaz cognitiva a través de la cual la gente puede entender o interpretar su ambiente social, lo que constituye el contexto de su discurso y de las prácticas sociales, entendido por Van Dijk como modelos de contexto, modelos mentales especiales de la memoria episódica que "controlan todos los niveles y aspectos de la producción y comprensión del discurso, tal como el género, las formas, el estilo, la variación y en general la manera en la que un discurso se adapta a la situación comunicativa." Van Dijk (Van Dijk, 1980).

Según este autor, los usuarios de la lengua adaptan sus producciones discursivas a lo que creen o saben que sus receptores ya conocen, es decir, adaptan su discurso al modelo del receptor a través, de lo que Van Dijk llama dispositivo de conocimiento, capaz de calcular en cada punto de la producción del discurso lo que los receptores ya saben por antecedentes y decidir qué conocimiento se debe presuponer, afirmar o recordar en determinado punto del discurso Van Dijk (Van Dijk, 1980).

La técnica de discurso explora el orden del texto y las formas en que el investigador organiza los textos, en este análisis se declaran los elementos sintácticos que exponen las relaciones entre oraciones y párrafos. Para ello se necesita estructurar el texto y describir las formas en que las locuciones y los párrafos se presentan, también es importante describir el orden en que se repiten y aparecen las expresiones de carácter nominal y sintagmático. Está técnica consiste en agrupar y analizar los textos que el usuario consulta y redacta. Estos textos son sometidos a la técnica de análisis contextual, que consiste en la detección de las unidades de relación del discurso, lo que

posibilita determinar el campo de acción y los nexos contextuales que tiene determinado usuario. De esta forma se pueden construir bases de conocimiento con los modelos de acción de los usuarios. En dichas herramientas quedará modelado su estilo de comunicación y de interacción con los saberes que se pretende resumir. El procedimiento propicia que cada usuario tenga un modelo de su discurso individual, algo vital para utilizar esta herramienta en la extracción del texto. Mediante la técnica de discurso es posible obtener el desarrollo de estructuras textuales de superficie y las llamadas macroestructuras textuales.

A diferencia de otros modelos de estudio de usuario, en este se utilizan textos generados por los mismos usuarios, lo que permite conocer los contextos donde se desenvuelven estos. La Técnica aplicada en este caso debe llevar asociados Mapas de Representación Conceptual los cuales se conforman con los segmentos temáticos que se intenta representar (ver ejemplo figura 14).


Figura 14. Mapa semántico de la Información Buzán (2003)

Esta técnica consiste desarrollar una organización abierta del contenido del texto que el usuario utiliza y genera, lo que posibilita una lectura que permite interpretar diversos elementos de la realidad del usuario, no expresado muchas veces en las investigaciones formales. La técnica se aplica de la siguiente

forma:

- Identificar la idea o las ideas principales del texto.
- Establecer categorías secundarias. (organizadas gráficamente en torno a la idea central).
- Determinación de los detalles complementarios.

El AD (Análisis de Discurso) no es la mera comprensión de lo que está explícito en el discurso, es un proceder mucho más poderoso. Según *Van Dijk* (Van Dijk, 1978) la comprensión es un proceso activo que no consiste solo en detectar las ideas que contiene el discurso y establecer la coherencia local entre ellas, sino en extraer el significado global –identificable, en cierto modo, con lo que se denomina tema– que posee y que va más allá de la suma de las ideas moleculares (microestructura) que lo constituyen.

Para Zaldua (Zaldua, 2006) el Análisis de Discurso (AD) realiza la interpretación y transformación de la información de un texto. Este instrumento académico provee al sistema de procesamiento de información de estrategias para percibir diversos elementos en el discurso mediante la inferencia de significados, esto posibilita la extracción de componentes que no han sido debidamente explicitados por el usuario. Para Zaldua (Zaldua, 2006) el AD no solo representa la información que esté en un discurso; el hecho de considerar la diferencia de dominios, la contextualización y lograr inferencias de contenidos implícitos en el discurso, avala a esta herramienta como útil para representar, además de información, conocimientos.

#### 4.3.5.- Etapa 2. Análisis manual del Corpus Textual

##### 4.3.5.1.- Fase 1. Análisis del corpus textual (representación del discurso en forma manual).

**Rhetorical Structure Theory tradicional (Mann & Thompson, 1988)**

Este proceso no es más que el análisis del corpus textual que ha de resumirse. Para realizar el análisis de discurso desde la perspectiva que se propone se tomará como sustento teórico metodológico la Rhetorical Structure Theory (RST) de Mann y Thompson (Mann and Thompson, 1990), creada al calor de

los estudios de generación automática de textos por un grupo de investigadores de la Universidad del Sur de California y ante la ineficiencia de modelos teóricos clásicos como los de Halliday y Hasan, Labov y Waletsky y Searle (Halliday and Hasan, 1976, Labov and Waletsky, 1967, Searle, 1975, Searle, 1969, Searle, 1965). La RST según D'Cunha (D'Cunha, 2006), tiene validez en la actualidad como una teoría descriptiva de organización del texto, caracterizando su estructura a partir de las relaciones que mantienen entre sí los elementos discursivos del mismo.

Esta teoría se nutre de una serie de enunciaciones, donde se aprecia la supremacía de estructuras con patrones de núcleo-satélites, la funcionalidad de la jerarquía y el rol comunicativo de la estructura del texto. Estos enunciados establecen registros de relaciones internas del texto, en las que algunas de sus unidades (satélites) aportan mayor especificidad que otras.

Según D´Cunha (D'Cunha, 2006) los 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.

Con frecuencia se observa que en el texto científico subyacen dos unidades de texto, aunque hay ocasiones en que es inevitable la presencia de nuevas formas de relaciones que poseen un rol determinado y que denotan la diferencia de roles relacionales en el ámbito de la RST. La RST establece en estos tipos de textos relaciones de elaboración entre las dos unidades adyacentes del texto D´Cunha (D´Cunha, 2006). Dicha relación también instituye que la afirmación es preponderante en el texto y menos usual en los segmentos de información adicional, por tanto la afirmación se erige como el núcleo de la relación y la información adicional en su satélite. No existen regulaciones discursivas incondicionales relacionadas con las unidades núcleo y satélites del texto, aunque debe apreciarse que la mayor parte de las relaciones pueden encontrarse de la forma en que estime el analizador del texto. En estas formas de construcción existen, relaciones similares entre las que pueden encontrarse: Evidencia, Fondo, Preparación o Concesión (Ver Anexo 10). Este análisis se lleva a cabo realizando marcas de discurso y de segmentos textuales. Mediante este estudio de análisis retórico se determinan

los marcadores del discurso y las diversas unidades elementales que se definen en el esquema del texto (Ver Anexo 10).

En ocasiones las relaciones que no presentan una unidad central con respecto a los propósitos del autor, se denomina Multinucleares.

En el gráfico (ver figura 15) observamos un ejemplo de un fragmento de estructura arbórea con relaciones de la RST, donde se ofrece una relación de Concesión, otra de Elaboración y una última relación Multinuclear de Lista:

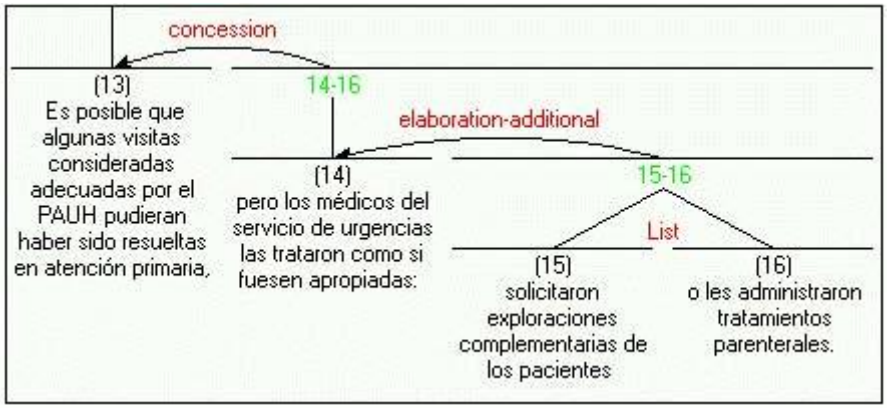



Figura 15. Fragmento de estructura arbórea con relaciones de la RST para artículos médicos
D´cunha, (2006)

Marcu (Marcu, 1997), formula para la RST sofisticados análisis que facilitan la segmentación del texto en unidades mínimas y propician que el conjunto de relaciones que se sustentan entre ellas proporcione la formalización de la estructura retórica arbórea mediante la división correcta y exacta entre los elementos núcleo y los satélites, los cuales se asocian a las relaciones discursivas, con una orientación hacia la generación automática de resúmenes. Marcu (Marcu, 1997), propone además un prototipo de analizador discursivo automático, que se basa en gran medida en el uso de marcadores discursivos. Para realizar este análisis es necesario contar con la herramienta Browser Versión tres y guardar los documentos y sus segmentos anotados en la base de datos o analizar el corpus manual y realizar la misma operación. La forma manual de aplicación de esta técnica se hace a través del análisis de los segmentos textuales y determinando las relaciones del texto (Ver Anexo 10) para ello solo basta con analizar qué es lo que declara el texto y clasificarlo.

#### 4.3.5.2.- Fase 2. Análisis sintáctico de dependencias

En los trabajos de Hays (Hays, 1960, 1964) es donde aparece por primera vez el vocablo sintaxis de dependencias, aunque más tarde Robinson (Robinson, 1970) también usara este término como gramática de dependencias. Gracias a las investigaciones de Tesnière (Tesnière, 1959) este proceder se erige como una representación convencional de la estructura sintáctica de las oraciones. La sintaxis de constituyentes, acuñada con anterioridad por Bloomfield (1933) citado por D´Cunha (D'Cunha, 2006), fue ganando terreno y desplazó a la de dependencias, relegándola a un segundo plano. Esta popularidad en el uso de este proceder se debe a que Chomsky (Chomsky, 1965) comenzó a utilizarla en sus trabajos que son la base del movimiento Generativo-Transformacional que continúa con Marvin Minsky.

Mediante la sintaxis profunda de dependencias asociada a la Meaning-Text Theory Mel'cuk (Mel'cuk, 2003), pueden confeccionarse estructuras basadas en árboles. Este proceder consiste en analizar nuestro corpus de artículos científicos y reducirlo a un grafo de orden parcial en forma de estructura de nodos donde se declaren las dependencias oracionales, 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).

Además es vital tener en cuenta la contraposición existente entre Tema y Rema, información importante para el resumen, pues estos elementos en la estructura comunicativa profunda integrada en la Meaning-Text Theory serán de importancia en casos en que la simple utilización de ella no pueda ofrecernos información suficiente para realizar generalizaciones según apuntan Mel'cuk y D'Cunha (Mel'cuk, 1988, D'Cunha, 2006). La herramienta de la que nos vamos a servir se llama IHMC Cmap Tools, desarrollada específicamente para la concepción de mapas conceptuales, pero que nos permite reflejar con claridad cualquier tipo de elemento o relación deseada. Esta herramienta permite marcar todos los casos de co - referencia. IHMC Cmap Tools no es la única herramienta que puede hacer tal actividad, existen otras cuyo valor en el tratamiento del texto es innegable.

#### 4.3.5.3.- Fase 3. Declaración de las reglas sintácticas de dependencia

La sintaxis de dependencias, que se integra en la Meaning-Text Theory presupone que un Modelo Sentido-Texto es un dispositivo lógico o conjunto de reglas que tiene como estructura general que permite establecer relaciones entre la semántica (significados estructurales), sintaxis (estructuras conectivas), morfología (forma en que se estructura el texto), las cuales son detalladas a un nivel de expresión mayor (Mel'cuk, 1988). Gracias a este proceder se describe la estructura del texto y sus relaciones.

Para su aplicación realice lo siguiente:

- Seleccionar el texto que va a resumir
- Identificar los párrafos que ha de segmentar y declarar.
- Analizar las oraciones para declarar las relaciones semánticas en forma de cadenas y seleccionar las palabras que se van a analizar.
- Considerar la naturaleza sintáctica de las oraciones para declarar las funciones que tienen los términos que aparecen en ellas.
- Declarar la forma en que aparecen los términos esenciales de la actividad que se trata.
- Identificar los elementos que tienen diferente proyección fonemática, esto evitará problemas en la comprensión. (Ver Figura 16).

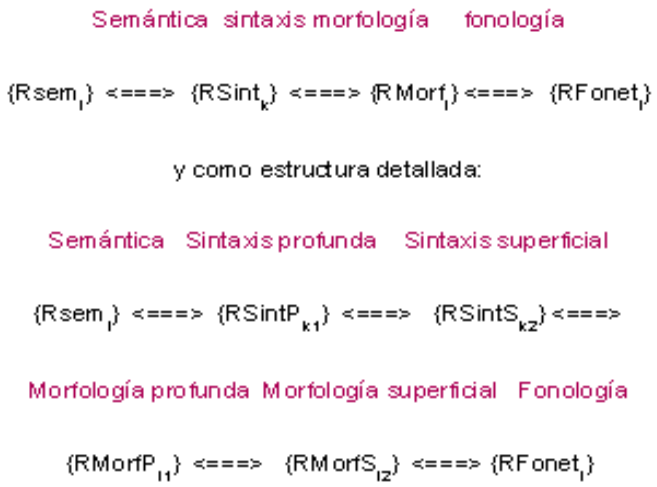



Figura 16. Declaración de las relaciones entre términos (D´cunha, 2006)

#### 4.3.5.4.- Fase 4. Desarrollo de las reglas sintáctico-discursivas

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) D´cunha (D´cunha, 2006). Estas reglas facilitan establecer la sintaxis entre las unidades del discurso (Ver ejemplo 1) y (Ver Anexo 7 y 10).

**Ejemplo 1: Reglas Sintáctico Discursivas (D´Cunha, 2006)**

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 se observa un satélite de una relación de Elaboración que además sea un ATTR, debe ser eliminado. Ocasionalmente estos segmentos son grupos de 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 78,8% de Polietileno, que emplean, también, en el servicio de Sistemas de construcción de enlaces de plástico. (m00788)[8] (Ver Anexo 7 y 10).

Después de analizar los textos que se van a incluir en nuestro sistema y todas sus relaciones es entonces que se debe confeccionar la base de conocimientos. Los procesos anteriormente descritos facilitan que los segmentos textuales que se integran a la base de conocimientos tengan la definición requerida y que el análisis a nivel macroestructural y microestructural permita utilizarlos en los procesos extractivos. También se logra mediante esta fase mostrar qué secciones del texto son preferibles para el resumen, lo cual representa un análisis retórico de elevado nivel de complejidad que garantiza la calidad de los procesos extractivos futuros.

### 4.3.6.- Etapa 3. Creación de la Ontología

#### 4.3.6.1.- Algunos Modelos de Ontologías

En esta investigación se describen algunos modelos de ontologías, se reconoce que existen diversas formas ontológicas (ver Anexo 8), sin embargo se optó por desarrollar aquellos modelos de más frecuente uso en la literatura y además con determinada relación con los objetivos de esta investigación. Los modelos utilizados para desarrollar la ontología de este metamodelo son los siguientes: Wordned, EurowordNed y PuertoTerm, este último no es una ontología, pero su estructura conceptual permite construir modelos de tratamiento de texto con alto nivel de sofisticación.

##### *4.3.6.1.1.- WordNed: estructura inicial de EuroWordNed*

Las bases de representación léxica constituyen un elemento esencial en los sistemas de resumen pues a partir de estas se forman los llamados lexicones o sistemas ontológicos que poseen determinadas relaciones. Uno de los sistemas de presentación léxica más importantes que conocemos es Word WordNet, descrito por Miller (Miller, 1993), el cual merece especial atención por varias razones. En primer lugar, aunque su corpus procedimental está formulado únicamente para el inglés, es innegable que ofrece numerosos puntos de análisis para el sistema que se propone en cuanto a fundamentos teóricos.

Cabe destacar, además, que fue uno de los primeros intentos serios de desarrollar un lexicón multipropósito en forma de aplicación informática. WordNet es un procedimiento electrónico que permite desarrollar asociaciones, esto se apoya específicamente en teorías psicolingüísticas relativas a la organización de la información léxica en la mente del hablante Miller (Miller, 1993). WordNet constituye un intento de reflejar el modelo de memoria léxica basado en redes semánticas propuesto por Collins y Quillian en un modelo lexicográfico de organización léxica.

Los objetivos esenciales de WordNet son los siguientes:

1. Validar las teorías psicolingüísticas sobre organización léxicas anteriormente mencionadas.

2. Aplicaciones múltiples en diversos estudios y herramientas que requieran acceso a información léxica.

Lo que distingue a WordNet de otros sistemas de implementación de lexicones computacionales es que es el único sistema de léxicos que es capaz de organizar el conocimiento sobre una estructura verdaderamente semántica y sobre la base de la estructura del lexicón mental. WordNet sigue la estructura de un diccionario conceptual, es decir se sale de los marcos de la estructura de un diccionario alfabético, para convertirse en una herramienta conceptual.

Si se observa un diccionario tradicional las diferencias son notables, pues WordNed divide el lexicón en cinco categorías: nombres, verbos, adjetivos, adverbios y elementos funcionales, sin embargo; WordNet elimina la información redundante que no aparecería en un diccionario tradicional, en aquellos casos en que una palabra pertenezca a más de una categoría. Este tipo de organización, permite el análisis de las disímiles formas de organización semántica que existen entre esas cinco categorías sintácticas, y también es importante destacar que, al no tener que forzar las diferentes categorías en un mismo esquema representacional, se puede buscar la forma más adecuada para cada una de ellas por separado.

La fundamentación teórica del sistema tiene su origen en la idea de la "matriz de vocabulario" ("vocabulary matrix") de Miller (Miller, 1993), proceder que usa el término *forma léxica* ("word form") para referirse a la expresión física que se escribe o se pronuncia y *significado léxico* ("word meaning") para referirse al concepto lexicalizado que se expresa por medio de una forma léxica. Este sistema propone la construcción de la escritura de los conceptos a través de matrices de vocabulario a los que se les denomina "synonym sets".

Esta Herramienta obviamente conlleva altos niveles de redundancia en cuanto a representación se refiere. WordNet se especializa esencialmente en la sinonimia, aunque no es la única forma de describir el léxico, también posee relaciones de hiponimia, meronimia, e hiperonimia y antonimia. WordNet está organizado de acuerdo a las relaciones semánticas, puesto que las mismas son relaciones de significados, y los significados están representados por medio de "synsets", WordNet expresa las relaciones semánticas como punteros

(*pointers*) entre "synsets".

#### *4.3.6.1.2.- EuroWordNet*

EuroWordNed es un proyecto para el Desarrollo de una Base de Datos Multilingüe WordNet con relaciones semánticas.

Los objetivos principales de EuroWodNed son los siguientes:

- Formular un sistema de bases de datos multilingües con vocabulario general caracterizado por la presencia de relaciones semánticas básicas entre palabras para distintas lenguas europeas (neerlandés, italiano, español).
- Conectar cada una de las lenguas (*wordnets*) con WordNet 1.5 de inglés americano y con un índice de significados (*Interlingual Index* o *ILI*). También posee una ontología común, mientras que las propiedades específicas de cada lengua se mantienen en los distintos *wordnets* individualmente.
- La base de datos emplea una aplicación de recuperación de información multilingüe, con el fin de mejorar la *llamada* (*recall*) en recuperación de información mediante la expansión de la(s) palabra(s) clave de un usuario a un conjunto más amplio de variantes y palabras relacionadas en cualquiera de las lenguas interconectadas.

El objeto esencial de WordNed es desarrollar redes semánticas de carácter multilingüe, para facilitar el acceso desde una lengua a otra. El proyecto no persigue servir para una aplicación específica, sino para todo tipo de aplicaciones dentro del PLN, la Recuperación de Información, la Inteligencia Artificial, etc. Esto propicia la construcción de mapas semánticos *multidimensionales,* ya que en un *wordnet* individual cada palabra se coloca en relación con el resto de las palabras en esa lengua y, en la base de datos multilingüe, estas redes relativas o mapas se conectan entre ellos a través de un índice de significados del inglés. Podemos decir que las características específicas de la base de datos EuroWordNet son:

- Cada *wordnet* muestra las relaciones semánticas en un sistema interno a la lengua, donde se especifican las diferencias culturales y lingüísticas de la comunidad a la que está dirigido el sistema.
- Contiene relaciones multilingües entre cada *wordnet* y los significados del inglés.
- Mediante las relaciones multilingües, es posible ir de una lengua a otra, lo cual hace posible comparar los distintos *wordnets* para descubrir inconsistencias y diferencias entre las distintas lenguas.
- Cada *wordnet* está relacionado con una ontología común independiente del lenguaje y con etiquetas de los campos de conocimiento en los que se usa ese significado. Estos pueden ser utilizados por el usuario para adaptar la base de datos multilingüe sin tener que acceder a cada uno de los *wordnets* de cada lengua.

Otros propósitos de EuroWordNet son:

- Integrar segmentos de la arquitectura de los léxicos semánticos y bases de conocimiento, necesarios para desarrollar sistemas de reconocimiento y comprensión automática del lenguaje.
- Ser una estrategia inicial para la elaboración de un léxico computacional para la Traducción Automática.
- Erigirse como herramientas de aprendizaje de lenguas, en las que las personas que aprendieran una lengua pudieran ojear el vocabulario de una lengua a través de la red a partir de una palabra conocida para ellos.
- Los correctores gramaticales y ortográficos pueden hacer uso de la información semántica para lograr reglas más precisas.
- El etiquetado automático de significados en corpus textuales.

El método global utilizado para construir la base de datos multilingüe EuroWordNet evidencia los pasos necesarios para la construcción de la ontología, proceso que puede resumirse de este modo:

- Reconocer y extraer relaciones semánticas de Diccionarios Electrónicos (*Machine Readable Dictionaries*).
- Sistematizar las redes.
- Unir los *synsets* a los *synsets* equivalentes del inglés.
- Utilizar la red en la base de datos de Novell con el fin de comparar los diferentes *wordnets*, aplicar y validar experimentos de análisis de ontologías que evidencien la existencia de nexos entre los conceptos de forma sistemática.
- Establecer y normalizar las conexiones entre los términos y los conceptos para que la base léxica pueda usarse con aplicación informática de cara a la Web.

#### *4.3.6.1.3.- Puerto Term*

Para construir la estructura conceptual de un dominio determinado se necesita esencialmente de la elaboración de jerarquías terminográficas, obtenidas mediante la selección y extracción de la información conceptual en repertorios especializados, diccionarios, tesauros, etc. Estas posturas sirven de base metodológica para el tratamiento terminológico de varias disciplinas que poseen nexos con la lingüística (terminología, lexicografía, traducción especializada) y con la Documentación (diseño y gestión de tesauros y ontologías).

En el caso de la terminología, desde finales de 1980 se ha generalizado el uso de textos especializados como fuente de información y localización de vocablos técnicos para alimentar bases de datos terminológicas. En la actualidad, y gracias al avance de las tecnologías de la información, estos corpus suelen encontrarse en formato electrónico Mcenery y Wilson (Mcenery and Wilson, 1996). La Documentación, por su parte, emplea esta misma técnica para la realización de herramientas propias de los lenguajes documentales, como los tesauros y, gracias al desarrollo de conceptos como la web semántica y las ontologías. Una vez más queda patente la estrecha relación que existe entre ambas áreas y que ha sido defendida por muchos autores como Irazazábal

(Irazazábal, 1996); Pérez Álvarez,(Pérez-Álvarez, 1998), Pinto y Cordón (Pinto and Cordón, 1999) y Sales (Sales-Salvador, 2006).

Si bien es cierto que la finalidad de ambas ha sido diferente (la terminología establece relaciones entre conceptos, y la Documentación se centra en la recuperación de la información), en esta tesis se ha intentado aunar ambas. Para llevar a cabo la estructuración del conocimiento se han empleado dos teorías lingüísticas con base semántica: el modelo lexemático-funcional de Martín-Mingorance descrito en Faber y Mairal-Usón, (Faber and Mairal-Usón, 1999) y en la semántica de marcos –*frame semantics*- también explicado en Fillmore, Fillmore y Atkins y en Gahl (Fillmore, 1982, Fillmore and Atkins, 1998, Gahl, 1998b, Gahl, 1998a).

El modelo lexemático-funcional facilita la representación de relaciones conceptuales y su posterior clasificación dentro de un lenguaje general y/o especializado.

Básicamente propone una distinción entre relaciones sintagmáticas y paradigmáticas –basada en los principios complementarios de combinación y selección Lyons (Lyons, 1977) que permite una organización conceptual independiente del sistema lingüístico. Ofrece un metamodelo lingüístico para la organización de conceptos. De hecho, concibe nuestro propio lexicón mental como una compleja red en la que cada nodo es un concepto y estos están interconectados por diferentes tipos de relaciones López-Rodríguez (López-Rodríguez et al., 2006). La semántica de marcos descrita en Fillmore, Jonson y Petruck (Fillmore et al., 2003), entiende el concepto de marco como una representación esquemática de un conjunto de conceptos relacionados entre sí.

El proyecto *FrameNet*, de la *Universidad de Berkeley* se basa en dicha teoría. Su gran ventaja radica en que la utilización de un único concepto activa todo el sistema conceptual permitiendo, gracias al empleo de enlaces web en su fase de implementación "física", conectar todos los conceptos subordinados y superordinados, facilitando la representación genérico-específica de los términos. Un desarrollo más profundo de esta fase del proyecto se puede encontrar en Faber (Faber et al., 2006).

#### 4.3.6.1.3.1.- Descripción de Procedimiento de Puerto Term

La primera fase se inicia empleando técnicas de recuperación de información desarrolladas en el ámbito de la lingüística de corpus, disciplina que ha mostrado su utilidad en proyectos a gran escala como las dos fases de *Acquilex* o, a menor escala dentro del mundo de los lenguajes documentales, los que pretenden generar clasificaciones, tesauros u ontologías.

Para Senso, et. al. (Senso et al., 2007) estas técnicas son muy útiles, ya que desde el punto de vista práctico sirven como una vía fácil para la obtención de amplios volúmenes asociados de terminología referente a un área determinada del conocimiento. Desde el punto de vista pragmático mediante el análisis de los vocablos, se asegura que la información extraída sea el reflejo tanto de los contenidos reales del dominio al que se destina el sistema, como del sublenguaje especializado empleado en el mismo.

Luego del estudio de los recursos terminológicos, según Senso, (Senso et al., 2007) se han creado otros nuevos, pues las disciplinas están en constante movimiento con una terminología en constante proceso de fijación y estandarización. Como es habitual, las fuentes de información empleadas para este análisis de corpus deben ser obras de referencia especializadas (enciclopedias y diccionarios técnicos mayoritariamente), monografías y artículos de publicaciones científicas. Estos documentos se almacenaron en ficheros de texto en diferentes directorios, dependiendo del idioma en el que estuviese expresado el texto, para facilitar su procesamiento por el resto de herramientas.

En esta fase, los términos recogidos forman una lista plana, donde ninguno tiene más valor semántico que otro. El siguiente paso se centra precisamente en formular un compendio de los conceptos más importantes. Para ello, y con la ayuda de especialistas y de la herramienta de análisis léxico Wordsmith Tools, distribuida por Oxford University Press y que permite explotar grandes conjuntos de textos mediante búsquedas basadas en parámetros contextuales o estadísticos, se obtiene una lista de frecuencias que permite inferir el conocimiento especializado en el campo de la ingeniería de puertos y buques. Mediante la identificación de las palabras clave se logra el modelado

conceptual de la ontología que articulará todo el conocimiento de esta disciplina.

Los lemas más frecuentes permiten identificar las categorías conceptuales sobre las que se fundamenta la definición de los términos del texto. Se deben usar varias técnicas de lematizado para conseguir una mayor adecuación en el empleo de cada término, y poder enmarcarlo dentro de su uso correcto. Mediante el uso de concordancias (la presentación de las ocurrencias de una palabra en su contexto lingüístico) se intenta buscar una nueva técnica que se asemeja según Senso, (Senso et al., 2007) a los procedimientos en los índices *KWIC* (*keyword in context*) pero cuya finalidad difiere de la indización permutada, pues la primera pretende indizar y recuperar documentos y la segunda contextualizar términos.

Se puede sintetizar la arquitectura de la ontología de un modo similar a WordNet al igual que las relaciones de términos. Algo que caracteriza a este sistema ontológico es la distinción entre clase e instancia, pues se refiere a la distinción entre significados denotativos y significados referenciales. Al igual que en WordNet, las meronimias se heredan a través de la hiponimia. Sin embargo, EuroWordNet posee un rasgo de disyunción y conjunción que permite representar dos típicos casos distintos de relaciones múltiples de meronimia: aquellos que se excluyen entre sí y aquellos que no. En el EuroWordNet, todas las relaciones poseen este rasgo. EurowordNet es un sistema de ontologías generales, por tanto no es un sistema apropiado para desarrollar una base de conocimiento especializado, si bien esta herramienta ha funcionado correctamente en sistemas de corte diccionario es imposible su uso en el resumen de textos especializados.

Para desarrollar esta herramienta léxica es necesario desarrollar una ontología que posea determinados agentes automáticos que sirvan de representación de los modelos cognitivos de los usuarios del sistema y por la que puedan ser descritos los cimientos inherentes a diversas materias. Acorde a las condiciones de nuestro país se debe usar el lenguaje de programación orientado a objeto. Esta ontología debe expresar las proposiciones lógicas del sistema, pues estas son las reglas bajo las cuales trabaja el mismo.

Es importante destacar que algunas de las relaciones que debe llevar la ontología fueron declaradas en procesos anteriores. Entre ellas se encuentran las reglas sintáctico-discursivas, que serán sometidas a verificación y a su posterior construcción en la referida ontología. Para la construcción de la ontología debe aplicarse un procedimiento sistemático y empírico de investigación tomando como precedente la construcción de un tesauro y el desarrollo de las teorías cognitivas de base. La ontología debe estar desarrollada a partir de los modelos cognitivos, lo que permitirá la recuperación de información, la recuperación de fragmentos de textos y el propio resumen.

Cuando se hizo el análisis de los corpus textuales pudo determinarse que la ontología debía contener alrededor de 7 600 conceptos donde se alojen alrededor de 9 600 propuestas cognitivas representadas en un lenguaje de definición. Estas propuestas están combinadas para recoger 7500 expresiones de contexto que se establecen entre las expresiones textuales y la pragmática.

Dichas expresiones representan los elementos del conocimiento que son demasiado grandes para una sola propuesta. Las propuestas estarán equipadas con descripciones semi-formalizadas que presentan formas superficiales de las propuestas que pueden asumirse en un texto. Entre los lenguajes más utilizados para hacer ontologías se encuentran OIL (Fensel, 2000) y DAML (Hendler, 2000). Estos lenguajes tienen una semántica bien definida y permiten la manipulación de taxonomías complejas y relaciones lógicas entre entidades en la Web.

Los referido lenguajes pueden “traducirse” al lenguaje RDF según Lassila y Swick (Lassila and Swick, 1999) (Resource Description Framework) mantenido por el consorcio W3C y a su complemento RDFS (Birkley, 2000). Actualmente existen editores que permiten generar el código de una ontología en diferentes formatos, como OILED o PROTEGE.

#### 4.3.6.2.- Pasos para confeccionar la ontología

Se reconoce que no existe un modelo de confección de ontología estandarizado, el autor ha decidido declarar la proyección especificada en el proyecto Puerto Term descrito en Senso (Senso, 2007, et. al.) de la Universidad de Granada por su flexibilidad y ergonomía. A continuación

presentaremos todo los elementos necesarios para la construcción de la ontología asociada a este modelo extractivo.

1. **Extracción de datos de diversas fuentes asociadas al dominio:** vocabularios controlados, corpus de sentencias, extracción en texto libre, preguntas de usuarios, etc.
2. **Declaración y Utilización de métodos para la extracción de conceptos:** análisis sintáctico, procesamiento del lenguaje natural, implicación humana, etc.
3. **Construcción y Utilización de métodos para la extracción de relaciones**: normalmente de forma automática, basándose en algoritmos.
4. **Reutilización de ontologías:** Localizar una ontología general asociada al dominio y reutilizar sus relaciones en función del usuario**.**
5. **Representación de la ontología:**
   - Declaración de la estructura jerárquica.
   - Formulación de las relaciones lógicas.
   - Construcción de los grafos conceptuales y el xml.
6. Seleccionar una herramienta o sistema asociados. Python.

**Herramientas de la Ontología**

Protegé: Aplicación amigable de modelado de conocimiento. El modelado está basado en torno a los principios básicos tales como clases, instancias, funciones, relaciones, etc. http://protege.open.stanford./index.html (Ver Anexo 8).

**Base de Datos de la Ontología**

La base de conocimientos ha sido descrita bajo las relaciones especificadas en el proyecto Puerto Term (Senso, et.al., 2007) (Ver Figura 17).

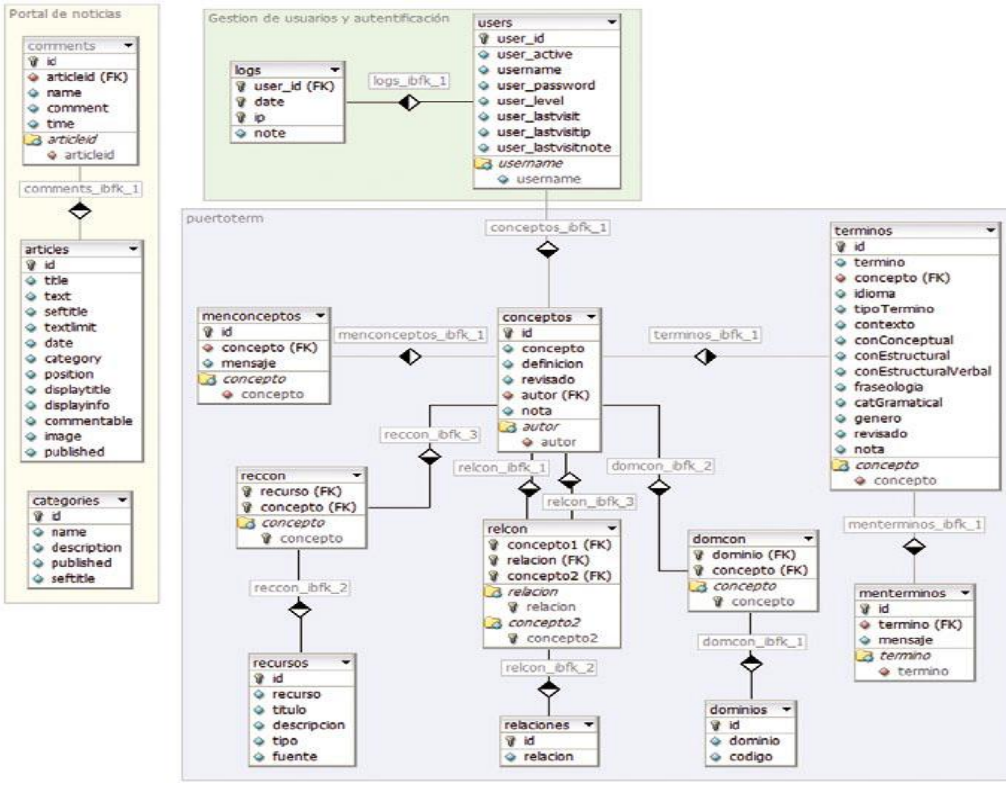



Figura 17. Base de Datos del Proyecto Puerto Term (Senso, et. al., 2007).

### *4.3.6.2.1.- Extracción de Texto con Agentes Cognitivos*

El proceso de resumen iguala la práctica humana competente como se ha descrito anteriormente. Se integra con actividades de búsqueda de información. Antes de que entren en función los agentes resumidores los documentos recuperados son reducidos a unidades del tamaño de un párrafo que contienen conceptos usados en la búsqueda. Los fragmentos resultantes son la entrada para el proceso de resumen y el equipo de agentes.

De acuerdo con los criterios de la comprensión humana, el equipo de agentes procesa uno por uno los fragmentos de texto que recibe. Algunos agentes se unen para decidir la relevancia de un elemento del texto. Revisan conceptos y su relación remitiéndose a argumentos típicos usados en el resumen (retórica). El conocimiento objetivo u otros conocimientos anteriores pueden influir en la calidad. Al terminar el proceso los agentes automáticos depositan los fragmentos de texto relevantes en la plataforma de respuesta a las preguntas (Ver Figura 18) con Diseño General del Sistema).

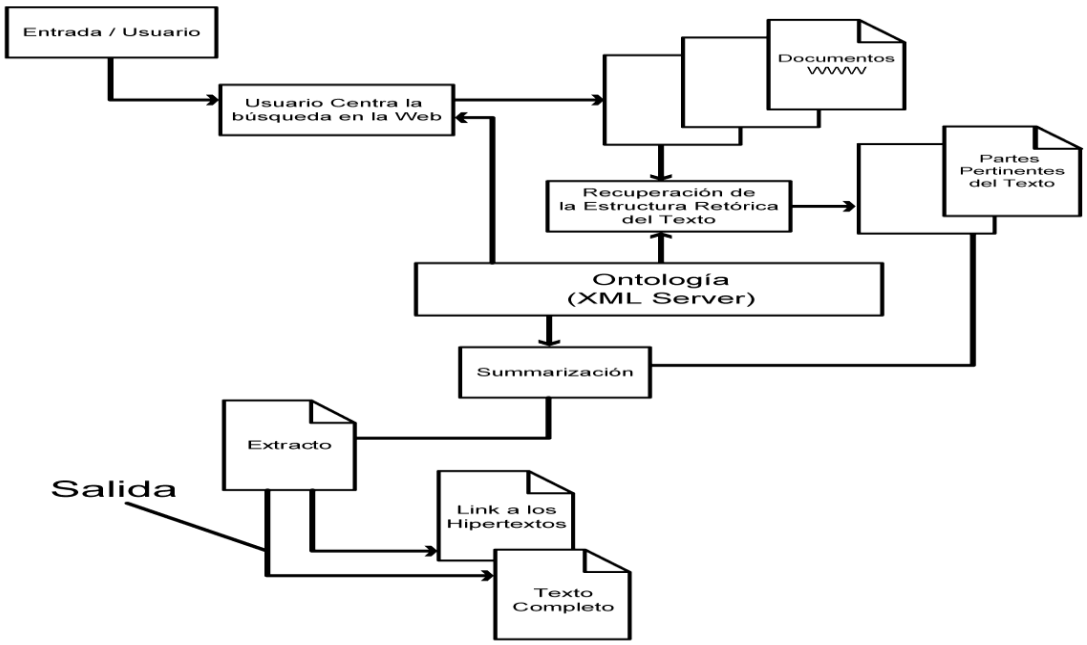



Figura 18. Arquitectura del Sistema

#### *4.3.6.2.2.- Extracción del Corpus Textual*

El corpus del texto se extrae mediante el contraste del lexicón o base de conocimientos contra el texto fuente. Para realizar la sumarización el sistema trabaja con diversas fases que se describen a continuación:

**Fase 1: Análisis léxico del texto:** Permite determinar las estructuras del texto y sus relaciones. Este procedimiento se hace mediante el concurso de software.

**Fase 2: Análisis sintáctico:** Mediante este se pueden analizar las relaciones que poseen los términos que se encuentran en el lexicón o sistema diccionario y someterlos a un análisis léxico, consistente en determinar los tipos de relaciones que deben haber sido declaradas de antemano en la formación del diccionario. Esto permite eliminar la ambigüedad. Para ello es indispensable el desarrollo de parser[26]inteligentes o sea agentes. Estos permiten identificar los elementos de las frases y determinar sus nexos. Otro elemento importante en el análisis sintáctico son los marcadores de cohesión que servirán de referente para el sistema de resolución de anáfora, pues ellos normalizan la relación de

[26] Analizador sintáctico automático

los elementos de una frase, es decir determinan las relaciones anafóricas, catafóricas, etc. (Ver Figura 19).

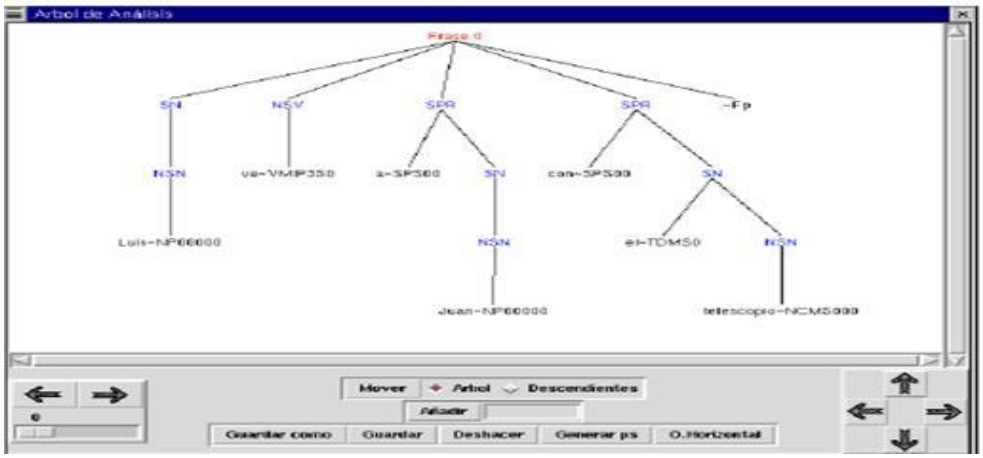



Figura 19. Analizador Sintáctico APOLN (Senso, 2009a)

**Fase 3: Análisis semántico:** La semántica en este modelo está ligada a los contextos donde se generan los contenidos que deben resumirse. El análisis semántico constituye un elemento importante dentro de la arquitectura del sistema, pues determina que significados tienen determinadas palabras dentro del texto. También incluye la construcción de un analizador semántico que facilite la agrupación de los significados y el sometimiento de estos a procesos de razonamiento.

**Fase 4: Desambiguación**

La desambiguación semántica (en inglés Word Sense Disambiguation, WSD) tiene como objetivo identificar los sentidos en que se expresa determinado vocablo. La DSA es denominada generalmente como desambiguación léxica (lexical disambiguation). Esta generalidad no es compartida por Yarowsky (2000), quien sostiene que la DSA toca sólo un segmento del vasto universo donde se encuentran las tareas de la solución de la ambigüedad léxica junto a la: generación texto-habla, clasificación de entidades discursivas y la corrección ortográfica.

Este método necesita un sistema de desambiguación con el cual, debe determinar los elementos que tienen nexo con los antecedentes y precedentes, es decir debe referenciar correctamente los elementos cohesivos del texto:

anáforas y catáforas. Mediante las relaciones de contexto y lo modelos de posición de sentido de palabras en el texto, puede mejorarse el problema que ocasiona la anáfora pronominal en lo referente a homogeneidad del texto. Una de las herramientas para las solución de problemas de desambiguación es STILUS, un software de análisis lingüístico de la firma DEDALUS S.A, quien se encarga de la revisión ortográfica, gramatical y de estilo de textos en español, no se describe una herramienta para los textos que están en ingles ya que el tratamiento de estos es menos complejo en este idioma, además Stilus corrige textos también en inglés. Todos sus procederes se realizan mediante complejos algoritmos de desambiguación léxica basados en el sentido de las palabras sobre la ontología.

STILUS permite la revisión de cualquier texto, sin importar su origen, además facilita la obtención de un informe donde se enuncian los errores ortográficos, gramaticales y de estilo detectados. Dentro de STILUS se encuentra el revisor ortográfico DEDALUS, el cual permite identificar en el texto, vocablos erróneos, desconocidos. Entre los errores más frecuentes que pueden detectarse mediante STILUS están los siguientes:

- Facilita la detección en los corpus de palabras erróneas o desconocidas y propone palabras alternativas.
- Dentro del sintagma nominal verifica la existencia de errores intrasintagmáticos, en caso que se produzcan violaciones de género y número.
- Fuera del Sintagma Nominal permite visualizar errores intersintagmáticos, producidos por violaciones de género o de número entre el sintagma nominal y los sintagmas relacionados a él.
- Detecta errores de secuenciación lógica, que violan restricciones de secuencialización de las categorías léxicas. (Ver ejemplo 2)

**Ej: 2**

Homofonía (p. ej., *Solo piensa en sí mismo.)(Villena et al., 2002).

Grupos verbales continuos (p. ej., *Había incluso pensado en venir antes).

§ Dequeísmo y queísmo (p. ej., *El servicio contestador de Telefónica le informa que no tiene ningún mensaje).

§ Sustitución de preposiciones (p. ej., *Los políticos discrepan con los sociólogos en muchas materias.).

§ Secuencias ilegales con amalgamas (p. ej., *Se trata de una sección rítmica a al que se añaden instrumentos.)"

STILUS también hace control de consistencia textual, para ello analizan dos niveles esenciales:

- Nivel léxico: En el nivel léxico, el control de estilo se efectúa sobre aquellos elementos textuales que impidan la legibilidad del texto
- Nivel estructural: en este nivel la revisión se realiza sobre el uso excesivo de determinadas construcciones lingüísticas, entre las que se encuentra la voz pasiva.
- Consistencia en el uso de convenciones de contenido: Controla los elementos convencionales del contenido del texto como las abreviaturas, los nombres propios, la escritura de fechas, siglas, acrónimos, etc.
- Los errores de puntuación más frecuentes en la redacción del texto dado por omisión o adición de un signo de puntuación, son elementos importantes en la revisión de estilo.

En el uso de ontologías en la desambiguación de texto se ha utilizado generalmente Eurowordned, pues posee redes semánticas estructuradas que facilitan la asimilación de posturas que marcan la relación entre palabras mediante el uso de funciones de similitud. Los primeros intentos de usar WordNet para DSA están declarados desde el campo de la Recuperación de la Información (Information Retrieval). Richardson y Smeaton (citado por Véronis, 200) generan una base de conocimiento para la jerarquía WN; aplicando una función de similaridad semántica descrita por Resnik para el terreno de la DSA y para IR. En Sussna (citado por Véronis, 2000) se propone una visión compleja del concepto distancia conceptual entre nodos (sentido de las palabras) partiendo del supuesto que un conjunto determinado de vocablos existente en un texto determinado los sentidos son elegidos mediante la

distancia mínima que existe entre cada uno de ellos. Si se quiere calcular la distancia semántica se asignan pesos a los accesos descritos por WordNet en concordancia con las relaciones de hiponimia, homonimia, hiperonimia, etc. y se establece una métrica que centra sus análisis en la cantidad de arcos que pertenecen a un mismo nodo y la profundidad del límite (edge) en el árbol general. Esta medida es empleada inmediatamente sobre los arcos del camino más corto donde dos nodos son unidos (sentido de las palabras). Los resultados de esta métrica estriban en un 55.8 % para vocablos o nombres polisémicos (71 %) a nivel de sentido, lo que indica que son mejores los últimos que la casualidad simple. Otro método empleado para el trabajo de la DSA es el procedimiento de Sussna a que es uno de los que hasta el momento utiliza algo más que la jerarquía es-un, utilizando otras relaciones en WordNet. Los trabajos de Resnik (1995) y Véronis (2000) se sustentan en el uso de una métrica de similitud semántica para palabras de la jerarquía del WN (Resnik 1993, 1995), calculando el contenido de información compartida por las palabras en la jerarquía es-un de WordNet. Este mismo autor Resnik (1995) confronta su técnica con los de otros autores que analizan la extensión del camino y discute y sostiene que los links en la jerarquía de WN no representan distancias uniformes. Sin embargo, su método que se basa en la granularidad final de WN, consigue un nivel de calidad (accuracy) análogo al de los humanos. Uno de los trabajos más actualizados es el de Rigau et al (1997) de se usan la medida de distancia conceptual, descrita en (Agirre et al. 1998), para evaluar la desambiguación semántica del genus en DGILE (Diccionario General Ilustrado de la Lengua Española) y en el LPPL (Le Plus Petit Larousse). Los resultados obtenidos mediante la aplicación de esta técnica reportan un 49% de precisión en el DGILE para nombres polisémicos (57% overall). Otra medida más compleja para medir la DSA es densidad conceptual desarrollada por Aguirre y Rigau a partir del estudio de la proximidad.

Basados en la capacidad operativa de Stilus como herramienta de desambiguación y los poderes semánticos construidos con la ontología es posible implementar en el modelo el algoritmo Aguirre (Aguirre, 1998) y Rigau (Rigau, 2002) el cual funciona mediante:

- Determinación de la totalidad de los nodos que identifican los sentidos de los vocablos a desambiguar W y los términos contextuales.
- Medir la densidad conceptual referente al sentido de cada término.
- Se determina el sentido con mayor densidad.

Se reconoce que existen otros métodos de desambiguación de probada calidad (entre ellos resolución de anáfora, etc.) que han surgido ante la inexistencias de ontologías disponibles para diversas actividades, pero en nuestro caso se decide aprovechar la ontología del sistema, ya que los resultados obtenidos con un sistema supervisado, nunca serán peores que los de un sistema no supervisado. Aquí se ha partido del sentído donde el texto como unidad integra debe estar desambiguado por eso un agente debe ejecutar la desambiguación de las palabras en la ontología, asignar reglas de puntuación al texto para su fácil lectura y ayudar en el etiquetaje de las palabras en el texto. La desambiguación culmina cuando el agente desarrolla el proceso de resolución de anáfora pronominal para dar cohesión al texto, que no solo será buscada mediante las marcas de discurso, sin que sea optimizada por la resolución de anáfora. El autómata trabaja sobre una modificación del algoritmo de MARS introducida por Domínguez (Domínguez 2011b) y Zidorov y Oliva (Sidorov and Oliva, n.d.) que en principio se ha desarrollado para el inglés por Mitkov (1998), el cual opera de la siguiente forma:

- En la *fase 1*, el texto a procesar se parsea a nivel sintáctico identificando todos los elementos que integran la estructura sintáctica de la oración entre los que se encuentran: la forma de los lemas, la sintaxis de las palabras, el número gramatical y las relaciones cohesivas del texto en términos de RST. Todo esto se hace con ayuda de la ontología y de la herramienta de Domínguez (Domínguez, 2011b).
- En la *fase 2*, Con Ontosatcol se identifican los pronombres anafóricos en plural y singular que demuestren anáfora nominal de identidad de referencia.
- En la *fase 3*, Cada pronombre anafórico es el antecedente de la parte que inicializa (párrafo, sección u oración). Seguidamente se busca los

antecedentes potenciales más cercanos a ese pronombre anafórico limitándose al párrafo del resumen.

- En la *fase 4*, el candidato oracional cuyo marcador sea el más alto es seleccionado como el antecedente del pronombre. Como el texto no es tan grande las uniones se formalizan seleccionando el candidato más reciente con el más alto marcador.

### 4.3.7.- Etapa 4. Modelación de Sistema de Búsqueda y Recuperación de información

La arquitectura de este modelo denota flexibilidad para el tratamiento de la búsqueda y recuperación de la información. Al desarrollarse una ontología que sirve de acceso a la información se hace necesario al menos delimitar algunos "slots" o ranuras como: publicado en, presentado en, temas y URL, lo que permite lograr una búsqueda de información coherente. Estos slots funcionan como anotadores de recursos. La búsqueda de recursos anotados es una nueva técnica de acceso a la información, y para explorarla debe construirse un software que pueda emplearse en cualquier servidor Web que soporte la especificación. La principal característica de nuestro buscador basado en ontologías reside en que los usuarios no necesitan introducir cadenas de caracteres que sirvan de patrón. Los criterios de búsqueda están previamente definidos en las ontologías, ya que se corresponden con las clases y subclases que los contienen.

El proceso de recuperación de recursos debe llevarse a cabo por tres caminos:

1. **Selección directa de Recursos Anotados:** Seleccionando directamente aquellos recursos anotados con los términos especificados en la interfaz, lo cual se traduce en recuperar las instancias de las clases seleccionadas. Nótese que si se especifica más de una clase, solo se recuperarán las instancias que pertenezcan a todas ellas (búsqueda *multicriterio conjuntiva*), no la unión de las instancias de cada clase.
2. **Recorrido de Relaciones Semánticas:** Recorriendo las relaciones semánticas establecidas entre instancias pertenecientes a alguna de las clases especificadas como criterio en la interfaz.

3. **Uso de las anotaciones:** la formulación de un criterio de anotación llevará al usuario directamente al recurso ya que la formulación del axioma incluye un tratamiento individual para cada elemento (ver Anexo 8).

Lo primero que tiene que hacer el usuario para activar nuestra futura herramienta de recuperación de recursos anotados es seleccionar los criterios de alto nivel que nosotros denominamos "puntos de entrada". Consideramos puntos de entrada a los términos más generales y significativos de la ontología que deben mostrarse en el buscador con el fin de hacer una primera acotación de la búsqueda. Obviamente debe existir la posibilidad de combinar más de un criterio para realizar la exploración. Dado que la ontología definida para artículos científicos, los puntos de entrada definidos en ella son casi todas las clases base, incluyendo, recursos *on-line* y publicaciones.

Una vez seleccionados los temas, se pueden buscar recursos o refinar la búsqueda (véanse los botones en la parte inferior de la interfaz). Los posibles caminos de navegación se muestran en el esquema (Ver figura 20).

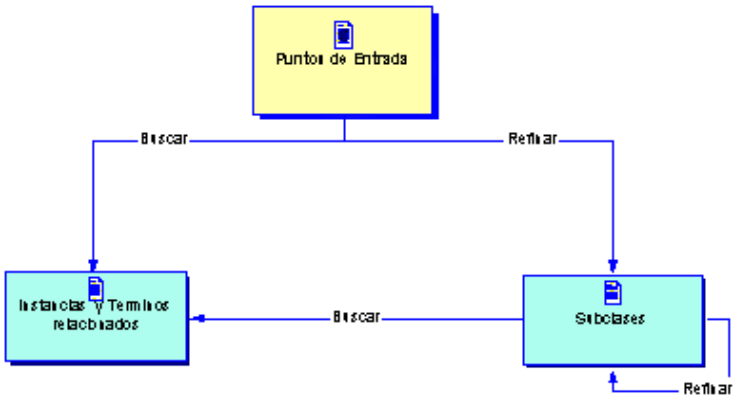



Figura 20. Propuesta de Esquema navegacional del buscador de recursos anotados (Estévez, 2009)

Para refinar los criterios de búsqueda, se muestran las subclases de las clases representadas por los términos seleccionados en la interfaz anterior. La posibilidad de refinamiento se sucede hasta que no existen más combinaciones de profundización en la jerarquía de los elementos que se van seleccionando.

Llegado este punto, la única opción disponible es buscar los recursos que respondan a los criterios especificados.

La opción *Buscar* debe dar como resultado lo siguiente: los recursos anotados como herencia de clase de todos los términos seleccionados previamente (semántica de conjunción) o de los términos correspondientes a subclases de los términos seleccionados y las instancias relacionadas. El potencial de la búsqueda específica de recursos anotados con términos de una ontología reside en la posibilidad que debe brindar el sistema para explorar relaciones semánticas entre las clases que la componen.

Esta forma permite la formulación de búsquedas multicriterio recorriendo los axiomas definidos entre los términos de la ontología y las instancias, dándole a la búsqueda un nivel semántico tal que las relaciones de simetría, asimetría y transitividad se convierten en enunciaciones del lenguaje formalizado. Este tipo de recorrido semántico permite también extraer temas relacionados con los recursos encontrados, habilitando la posibilidad de explorar nuevas vías que no son las especificadas inicialmente (un ejemplo de los resultados de búsquedas con temas relacionados), pues las instancias son hiperenlaces que llevan a información ampliada sobre las instancias o a nuevas búsquedas.

Para ilustrar un ejemplo de búsqueda veamos los pasos que se tendrían que realizar para encontrar artículos de revistas que versen sobre Electrónica Digital. En primer lugar hay que indicar los temas sobre los que se quiere buscar en la interfaz principal de la aplicación. Estos temas están declarados en la ontología. Al ser interés de los usuarios únicamente las publicaciones en revistas no hay que continuar refinando la búsqueda, el usuario puede hacer una búsqueda a partir de la subclase revista y encontrar allí todas las instancias de ese tipo de documento, por tanto el proceso de refinado no es un proceso clave en este sistema.

Esta acción visualiza las subclases de las clases especificadas dentro de nuestra ontología. Si dentro de los artículos solo queremos obtener los publicados en revistas, el solo tiene que decir introducir el axioma publicado en y tendrá todos los recursos anotados bajo este axioma.

### 4.3.8.- Recursos necesarios para el desarrollo del metamodelo

#### Agentes Automáticos

En el modelo, un grupo de agentes, que en realidad son estrategias cognitivas, se reproducen bajo modelos de procesamiento humano. En principio cada agente debe ejecutar una estrategia de resumen usada por seres humanos especializados en el procesamiento de la información. Estos pueden necesitar subagentes especializados para lograr todo el proceso. Todos los agentes en el sistema se implementarán como clases de Phyton.

Como Python está orientado hacia objetos, los agentes heredan de su clase superior de agentes algunos métodos fundamentales como la segmentación de documentos si se necesita información específica usando bases de conocimiento., Ej. La ontología o los unificadores. Estos métodos brindan el conocimiento básico necesario para el procesamiento y el resumen del texto. Pero los agentes lo utilizan de diferentes maneras dependiendo de la tarea dentro del proceso de resumen.

Los agentes de resumen deben trabajar sobre una pizarra basada en XML que sirve como principal medio de comunicación. Su encabezamiento establece el contexto inmediato del usuario, lo que fue establecido en la fase de análisis de dominio. El fragmento considerado para resumir es almacenado y procesado una vez más.

#### Agentes de Contexto

Los agentes de contexto se comportan como lectores humanos, que verifican si el documento que se necesita se corresponde con las necesidades expresadas por el usuario, buscando marcas específicas entre las que se encuentran: conceptos, descripciones, etc. El contexto muestra los documentos encontrados durante la recuperación de fragmentos de los conceptos del contexto elegido para el resumen. Estos conceptos no han sido declarados sistemáticamente en las encuestas de red. A diferencia de los agentes, los resumidores humanos saben cómo aprovechar la estructura de los documentos al igual que los agentes.

En artículos científicos, con una superestructura conocida limitan sus operaciones a las secciones llamadas “Introducción” y “métodos” o “Materiales y métodos”, conclusiones, etc. En otro tipo de documento no estructurado inspecciona todo el texto. Cuando el agente de contexto entra en funcionamiento todos los textos han sido convertidos a la estructura XML y así el agente puede navegar en ellos y no necesita conocimientos de otros formatos de documentos. Si este agente descubre al menos uno de los términos de interés, el documento se mantiene en su pizarra para futuros resúmenes potenciales, sino todo el documento es desechado y la pizarra es borrada.

**Agentes de Interpretación de Texto**

Los próximos cuatro agentes representan un modelo reducido de la comprensión humana y delimitan el tamaño del texto. Los agentes se software que han de implementarse son los siguientes:

- Proposiciones en el Texto: Dirige el texto hacia significados hipotéticos de cláusula u oración o sintagma verbal.
- Proposiciones Normalizativas: Los significados procesados por el agente de posiciones en el texto se unen a las proposiciones Normalizativas que ya tiene establecida la ontología. Aquí se verifica la jerarquía en que se ordena el conocimiento.
- Agentes Buscadores de Proposiciones: Se corresponde con una búsqueda de todo lo que significa conceptos de interés y otros elementos relacionados en el modelo de búsqueda propuesto en la ontología.
- Preguntas: Introduce la funcionabilidad para analizar conocimiento orientado. Se chequean las proposiciones buscadas en los segmentos de texto decretados para el resumen y se desechan aquellas que no están acordes con estas recomendaciones o que sencillamente no cualifican.

Basados en el conocimiento representado en la ontología y en sus Anexos, el grupo de agentes extractores es capaz de construir una representación interna

del texto. Se desechan todos los candidatos que no cualifican con la tarjeta o la retórica de sumarización y se reduce el material considerado, luego se declaran técnicas de interpretación de texto:

**La Posición en el texto.**

Acepta las oraciones de los párrafos candidatos y transforma entonces el texto basado en proposiciones preestablecidas.

**Normalización de la Posición**

El agente describe una posición preliminar proporcionada por el Agente de Posición en el texto ajustado a la posición que tiene en la base de conocimientos.

**Buscador de Proposiciones**

Encuentra proposiciones conducidas a la comprensión del texto y las lee como un lector profesional o especializado. Los conceptos de la ontología usados en una posición candidata se utilizan como palabras para buscar la interpretación hipotética del conocimiento almacenado en la base del conocimiento.

**Agentes de Preguntas**

Los agentes de preguntas chequean en detalle si la formulación está relacionada con el estudio de necesidades del usuario. Aceptan propuestas que han sido interpretadas exitosamente por el agente de búsqueda de propuestas y busca sus nexos en el contexto del las necesidades del consumidor.

**Agentes de Lectura de Imágenes**

Describen y leen los elementos gráficos que tiene el sistema y lo conectan a los textos seleccionados en la ontología, de acuerdo con descripciones connotativas.

**Agentes de Redundancia**

Busca las entidades que se repiten en las propuestas que quedan del resumen, pues las repeticiones estructurales son frecuentes en los textos científicos. Su propósito está orientado al lector, buscando temas anteriores antes de que sean vueltos a difundir o que aumente la información en ellos. Al buscar

material redundante los resumidores humanos reducen el tamaño de la información sin pérdida de la misma. Los agentes de redundancia siguen el ejemplo de los resumidores humanos, buscan información duplicada y la retiran.

### 4.4.8.- Representación Textual de la Información

Para representar la información nuestro modelo necesita cambios intrínsecos que obligan a utilizar técnicas de extracción de documentos múltiples. El resumen es texto y la única forma de representarlo es mediante técnicas de agrupamiento documental que estén descritas en diversas herramientas.

En el caso de nuestro metamodelo la herramienta utilizada es el agoritmo generado en Corpus Miner de Arco (Arco, 2007) mediante el cual se obtienen agrupaciones de términos que representan los contenidos de los documentos, y puede utilizarse como un sistema de clasificación. La flexibilidad con que fue realizado el sistema de Corpus Miner permite la descripción de los documentos si se conectaran sus módulos de programación al diccionario de la ontología. Antes de realizar esta investigación era solo una herramienta para resumir y agrupar textos.

Hoy se utiliza en el tratamiento de resúmenes de patentes y espera su aprobación por el CITMA (Ministerio de Ciencia Tecnología y Medio Ambiente) como Proyecto para el Desarrollo de la Información y la Tecnología. Los pasos para trabajar un sistema como Corpus Miner son los siguientes (ver figura 21 y 22) (ver Anexo 9 y 10).

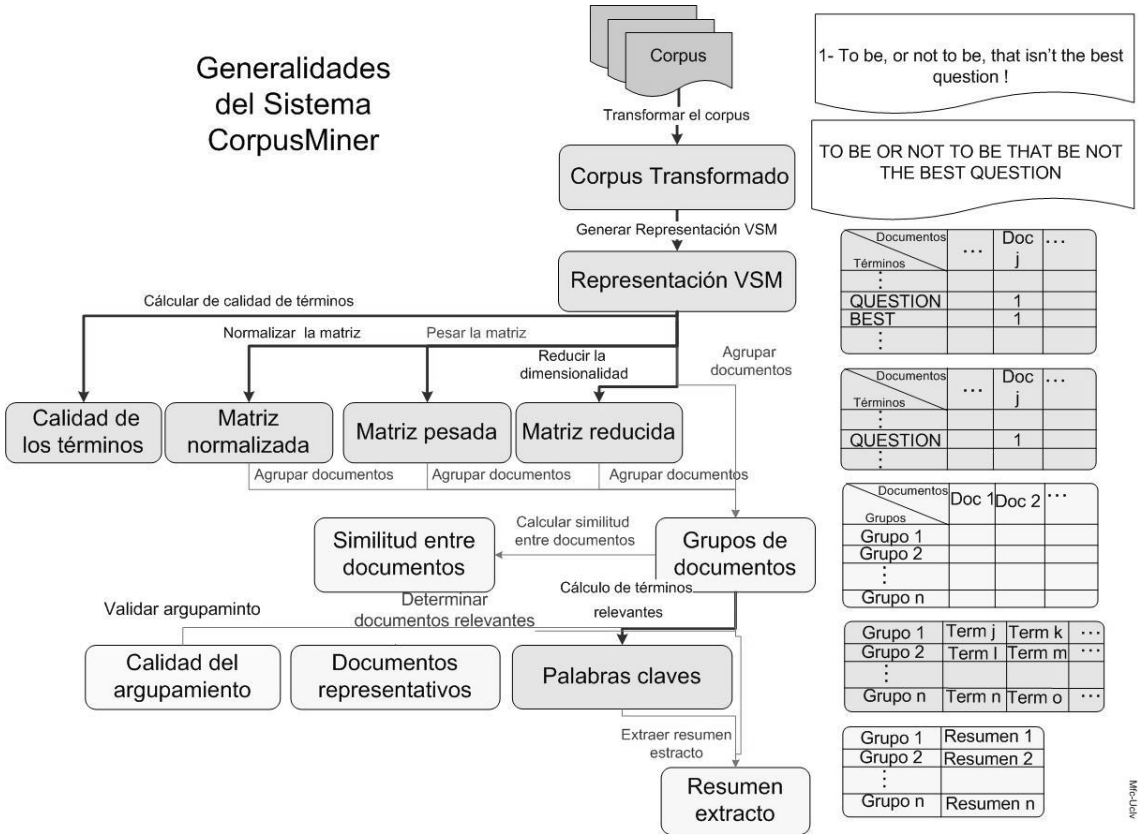



Figura 21. Modelo integral del sistema Corpus Miner, en Arco (2007)

### 4.4.8.1.- Fase 1 Representación Textual

Representación de la VSM (Vector Space Model).

| | Término 1 | Término 2 | … | Término *m* |
|---|---|---|---|---|
| Documento 1 | $tf_{d_1}(t_1)$ | $tf_{d_1}(t_2)$ | | $tf_{d_1}(t_m)$ |
| Documento 2 | $tf_{d_2}(t_1)$ | $tf_{d_2}(t_2)$ | | $tf_{d_2}(t_m)$ |
| … | | | … | |
| Documento *n* | $tf_{d_n}(t_1)$ | $tf_{d_n}(t_2)$ | | $tf_{d_n}(t_m)$ |

Figura 22. Representación del VSM

**Subfases**

- Transformar el corpus textual
- Normalizar y pesar la matriz

- Calcular las medidas de calidad de términos
- Reducir la dimensionalidad de la representación (Ver figuras 23 -28)

### 4.4.8.2.- Fase 2 Agrupamiento de Documentos

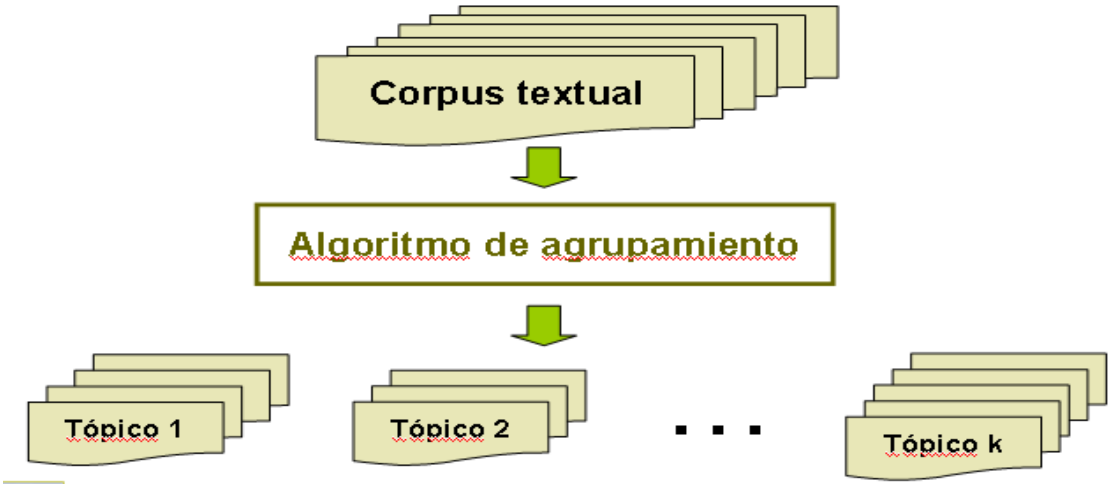



Figura 23. Agrupamiento de Documentos (Arco, 2007)

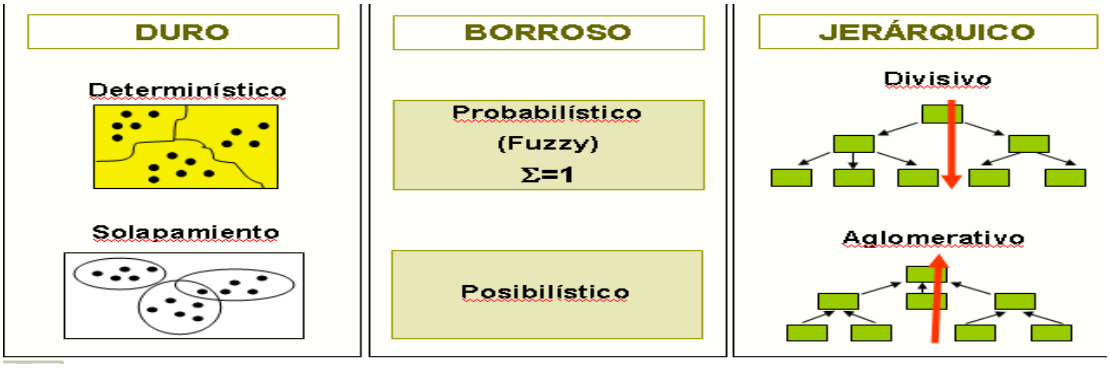



Figura 24. Técnicas Utilizadas en el agrupamiento (Arco, 2007)

### Algoritmos de agrupamiento en CorpusMiner

- Simultaneous Keyword Identification and Clustering of Text Documents (SKWIC) (técnica de análisis duro y determinista)
- Simultaneous Keyword Identification and Fuzzy Clustering of Text Documents (Fuzzy SKWIC) (técnica de análisis borroso)
- Extender Star (técnica de análisis duro y con solapamiento)

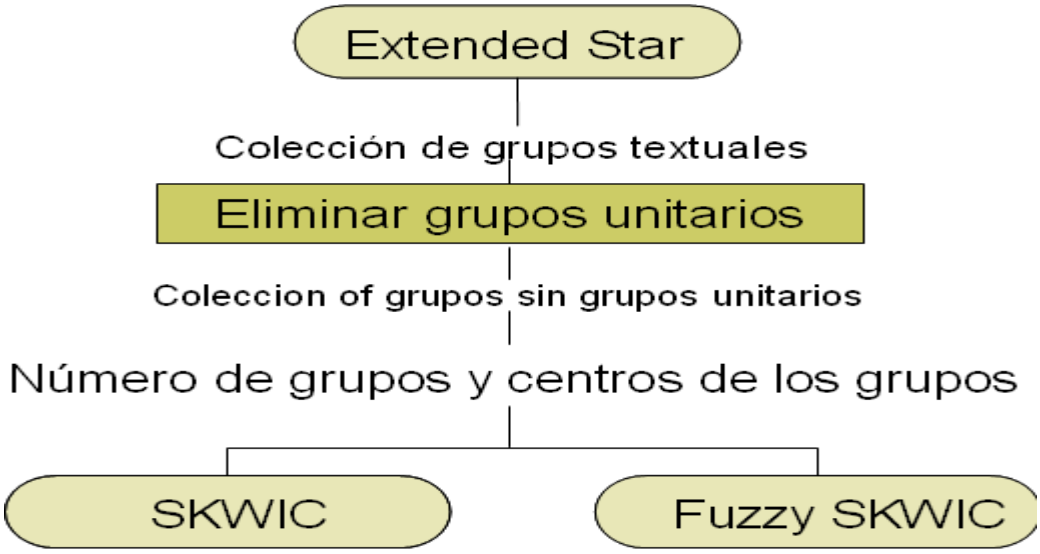



Figura 25. Métodos Concatenados de Análisis

**Subfase 3. Definición del Sistema de Decisión.**

Sistema de decisión $DS=(U, A\cup\{d\})$

$U=\{D1,D2,\ldots,Dn\}$ es la colección de documentos

$A$ es un conjunto finito de palabras clave que describe esa colección de documentos $d\notin A$ representa los resultados del agrupamiento (atributo de decisión) representa la frecuencia pesada del término $j$ en el documento $i$.

| | Término 1 | Término 2 | … | Término $m$ | Grupos |
|---|---|---|---|---|---|
| Documento 1 | $tf_{d_1}(t_1)$ | $tf_{d_1}(t_2)$ | | $tf_{d_1}(t_m)$ | $Grupo_{t1}$ |
| Documento 2 | $tf_{d_2}(t_1)$ | $tf_{d_2}(t_2)$ | | $tf_{d_2}(t_m)$ | $Grupo_{t2}$ |
| … | | | … | | … |
| Documento $n$ | $tf_{d_n}(t_1)$ | $tf_{d_n}(t_2)$ | | $tf_{d_n}(t_m)$ | $Grupo_{tk}$ |

Figura 26. Decisiones de Términos

### 4.4.8.3.- Fase 3 Extracción de palabras clave en grupos textuales homogéneos

- Selección de las palabras de mayor relevancia resultante de métodos de agrupamiento.

- Selección de los términos con mayores valores de calidad en el grupo.
- Selección de los términos a partir de las reglas generadas por el algoritmo ID3 en cualquiera de sus variantes.
- Variantes combinadas.

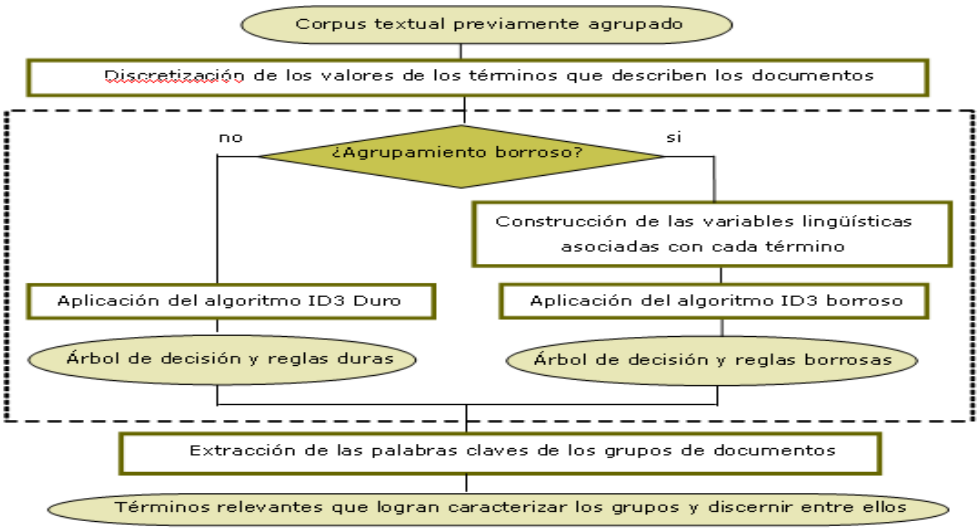



Figura 27. Resumen de la fase de Extracción de palabras clave en grupos textuales homogéneos Arco (2007)

**Subfase 1 Relación de similitud *R'***

- Sea $s:U\times U\rightarrow R$ una función que mide la similitud entre documentos de $U$.
- Utilizar, Jaccard y el coeficiente Coseno, pues son los más usados.
- $R'(x)=\{y\in U : yR'x,$ i.e. $y$ está relacionado con $x$ ssi $s(x,y)>\xi\}\xi$ is a similarity threshold.
- Tenemos que calcular $R'(x)$ para cada documento $x$ en $U$.

### 4.4.8.4.- Fase 4: Extracción de las aproximaciones superiores e inferiores

- La aproximación inferior de cada grupo textual incluye los documentos más representativos del grupo.
- La aproximación superior de cada grupo incluye todos los documentos del grupo y aquellos que tienen relación con ellos.

- Es posible extraer las oraciones que tienen presencia de las palabras clave obtenidas en el tercer módulo, tanto de las aproximaciones inferiores como superiores de cada grupo.

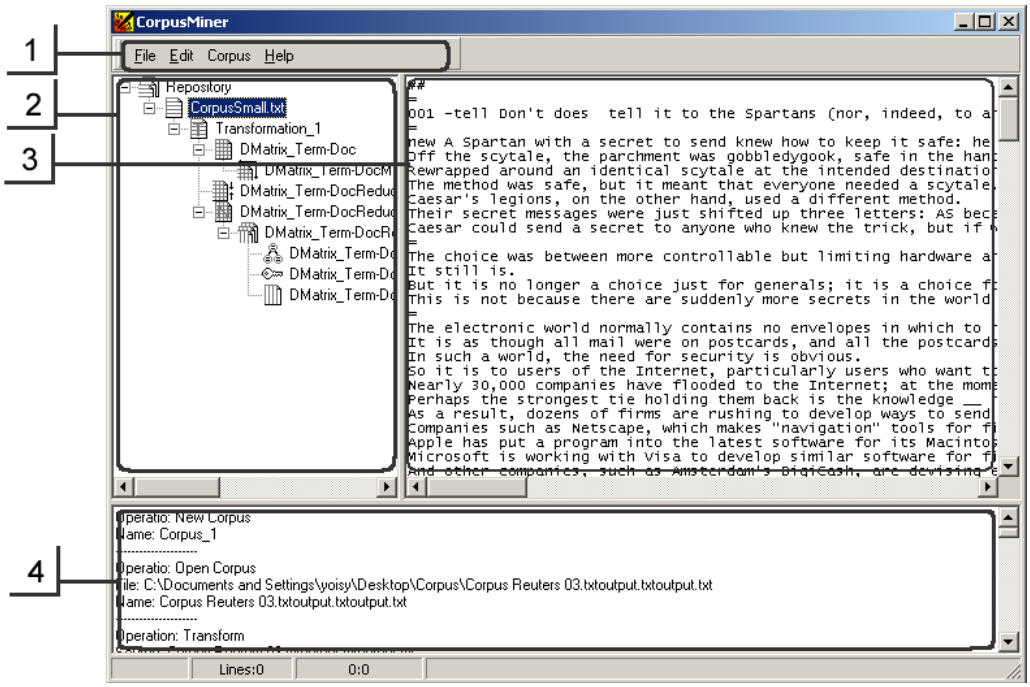



Figura 28. Herramienta Corpus Miner (Arco, 2007)

1. Menú principal.

2. Árbol de resultados.

3. Pantalla de resultados

### 4.4.8.5.- La Representación Textual en Corpus Miner (principales fases)

La representación textual es una tarea que permite corroborar la forma en que serán descritos los textos en el futuro modelo Lewis (Lewis, 1992). Para este Diploma de Estudios Avanzados se ha apelado a la variante (Vector Space Model; VSM) descrita por Salton, et al. (Salton et al., 1975) por su validez en el tratamiento de textos y en la representación documental y su amplia aceptación en el campo de la Minería de Texto. VSM es la representación que utiliza CorpusMiner, que es la herramienta de comprobación utilizada en esta investigación. Para el cálculo VSM según Arco (Arco, 2008) “se parte del supuesto de que cada documento es identificado como un vector de rasgos en un espacio en el cual cada dimensión corresponde a términos indexados

distintos (palabras)". En los trabajos de Lanquillon (Lanquillon, 2002) se expresa que un vector documento dado, en cada componente tiene un valor numérico. Para indicar su importancia se muestra que la transformación del corpus, la extracción de términos, la reducción de la dimensionalidad, y la normalización y pesado de la matriz, son elementos esenciales de la representación de términos.

**Transformación del corpus**

Cuando se dice que se quiere transformar el Corpus textual se intenta convertir los ítems de entradas (ficheros) en elementos lingüísticos (tokens de palabras). El paso que le sigue a la transformación del corpus es la extracción de términos, donde estos segmentos lingüísticos serán usados para generar rasgos significativos (índices de términos).

El primer paso en la trasformación del corpus es reconocer los componentes textuales desde los diferentes formatos, ya esto se ha realizado previamente pues los documentos en las ontologías están representados y transformados en txt, XML o HTML. En Corpus Miner existe el PDFBox45 para transformar los ficheros .pdf en formato texto. Esto permite que los documentos agrupados se conviertan en una secuencia de tokens.

Los Tokens son sometidos a diversas transformaciones entre las que se encuentra: convertir las letras todas a mayúsculas o todas a minúsculas, eliminar las marcas de puntuación al final de los tokens, omitir los tokens que contienen caracteres alfanuméricos, identificar o marcar los nombres de personas, localidades, organizaciones y productos, y sustituir las contracciones y abreviaturas por la expresión completa que representan Lanquillon (Lanquillon, 2002).

**Extracción de términos**

A partir de la secuencia de tokens, obtenida a partir de la transformación del corpus se produce una secuencia de términos indexados basados en esos tokens. En esta aplicación el vocabulario se crea a partir de los términos indexados resultantes de la extracción. Según Arco (Arco, 2008) representar documentos del lenguaje natural por el significado de un conjunto de índices de

términos es un reto, sobre todo porque la información siempre depende del contexto. Nótese que los niveles lingüísticos que se han declarado para el tratamiento de los textos son los siguientes y están basados en lo expresado por Lanquillon, (Lanquillon, 2002).

Nivel de grafema: Análisis sobre un nivel de sub-palabra, comúnmente concerniente a las letras.

Nivel léxico: Análisis concerniente a palabras individuales.

Los dos primeros niveles operan solamente con un plano estadístico sobre el texto, es decir básicamente sobre frecuencias de combinaciones de términos, que pueden ser letras o palabras.

3. Nivel sintáctico: Inspección concerniente a la estructura de las oraciones.

4. Nivel semántico: Estudio relativo al significado de palabras y frases.

5. Nivel pragmático: Se refiere al significado, tanto dependiente del contexto como independiente del contexto (por ejemplo, aplicaciones específicas, contextos).

En esta tesis el análisis léxico de los textos identificando las palabras simples como rasgos se basa en la descripción de Salton, y Buckley,(Salton and Buckley, 1988). Aquí solo se desarrolla la extracción estadística, pues la semántica fue hecha para conformar los resúmenes téngase en cuenta que ahora estamos representando en palabras claves aquellos documentos que ya fueron extraídos para que los usuarios puedan reconocerlos.

El modelo estadístico adoptado es la denominada bolsa de palabras o bag-of-words modelo descrito por Lewis y Ringuette (Lewis, 1992). Se selecciona este método debido a su capacidad de definir términos independiente del léxico o del domino que se trate, además su coste computacional es bajo, por lo que su rendimiento es muy eficiente, por otra parte las salidas de la representación son fáciles de interpretar por los usuarios. Según Arco (Arco, 2008) Los tokens son cadenas de caracteres delimitadas por espacios en blanco (por ejemplo espacios, cambios de líneas, tabs) con este tipo de rasgos es natural obtener el sistema de información requerido para la aplicación de RST.

La extracción de palabras independiente del dominio se considera un elemento ventajoso a diferencia del uso de frases que tienen a ser dependientes del dominio Sahami (Sahami et al., 1988) , esta afirmación es verdadera si no se quiere realizar o formular un experimento de una comunidad epistémica especifica y por consiguiente el tratamiento de las inflexiones y otras cualidades lingüísticas derivativas traerían problemas en las investigaciones, lo que demandaría de procesos de alta complejidad como la reducción de la dimensionalidad.

**Reducción de la dimensionalidad**

Este proceso constituye la determinación de la forma en que se hará más tratable el algoritmo. Según Arco (Arco, 2008) es esencial controlar la dimensionalidad del espacio del vector documento cuando se utilizan las palabras como los términos a indexar. Las causas que obligan a reducir la dimensionalidad son las siguientes según Arco (Arco, 2008):

La complejidad de muchos algoritmos de agrupamiento depende crucialmente del número de rasgos y reducirlo es necesario para hacer estos algoritmos tratables

Existen palabras que son irrelevantes y provocan la obtención de peores resultados, por tanto, eliminarlas puede realmente aumentar la eficiencia del agrupamiento a realizar.

En el trabajo de Arco (Arco, 2008) se reduce la dimensionalidad para abarcar técnicas que controlen la dimensionalidad del vector ellas son: selección de rasgos y reparametrización descritas por Lanquillon (Lanquillon, 2002), o la combinación de ambas técnicas según Nigam, (Nigam et al., 2000). La eliminación de palabras de parada o gramaticales denominada en el área de la computación (stop word elimination) utilizada por diversos autores como Yang y Pedersen (Yang and Pedersen, 1997) y Mladenic y Grobelnik (Mladenic and Grobelnik, 1998), es uno de los procesos selectivos que usa Arco García en su software.

Arco (Arco, 2008) utiliza ampliamente los métodos de filtrado para decidir la inclusión de un término en el vocabulario o no, tratando cada uno

independientemente y evaluándolo. El vocabulario resultante es desarrollado a partir de la selección de todos aquellos rasgos que posean una puntuación superior o inferior a un umbral predeterminado, o los mejores rasgos, es decir, los rasgos con mayor o menor puntuación acorde a la magnitud de la puntuación. En este software se han considerado los criterios siguientes para la evaluación de los términos:

Frecuencia de aparición en los documentos siguiendo la Ley de Zpif Cantidad de documentos en los cuales el término aparece Yang y Pedersen (Yang and Pedersen, 1997).

Elevada frecuencia de aparición respecto a la frecuencia inversa de documentos TFIDF Yang y Pedersen (Yang and Pedersen, 1997).

Las técnicas lingüísticas que reducen la dimensionalidad por reparametrización incluidas en esta aplicación son:

La homogeneidad ortográfica (spelling) proceso que facilita convertir todas las palabras del léxico en un idioma estándar.

La reducción de las palabras a su forma raíz (stemming) descrito por Yang; Pedersen (1997) que permite reducir la dimensionalidad del espacio de rasgos haciendo corresponder las palabras morfológicamente similares con la palabra raíz asociada Frakes, Baeza y Porter (Frakes and Baeza-Yates, 1992, Porter, 1980).

Otras técnicas como el análisis de latencia semántico (Latent Semantic Analysis) de Landauer, Dumais y Shaal (Landauer and Dumais, 1997, Schaal et al., 2005) no han sido incluidas por su alta complejidad computacional. No se proponen técnicas específicas para el tratamiento de ontologías y tesauros por ser esta una representación preliminar de lo que ya existe en una ontología.

El uso de herramientas de comprobación de ontologías está asociado a los trabajos de Shaal (Chal, 2005), y Meller, (Müller et al., 2005) será utilizado en los procesos de evaluación de la ontología.

**Normalización y pesado de la matriz**

De acuerdo las frecuencias estadísticas de los términos en los documentos, se genera un vector pesado para cualquier documento basado en el vector de frecuencias de términos. Cada ponderación refleja el valor de un término en un documento con respecto a su frecuencia en todos los documentos. Algunas variantes permiten pesar los términos indexados. En esta aplicación se ha reportado en las formas de pesado global alcanzadas variando la fórmula TF-IDF de Mannig y Shütze (Mannig and Shütze, 2000) y Berry (Berry, 2004) en los cuales la normalización se obtiene dividiendo la frecuencia de aparición de los términos por la longitud de los documentos, lo que permite solapar el posible tamaño no homogéneo de los textos.

**Descripción de Procesos**

Uno de los elementos que señala Moreiro (Moreiro, 2004) para el diseño de los sistemas de información, son los estudios de ingeniería de software. Estos permiten reconocer e interpretar las condiciones que tendrá el futuro software (ver figura 28 y tabla 10). En este acápite de la investigación se describirán algunos elementos de ingeniería de software que enuncian los roles que un usuario debe presentar ante el sistema.

| **Qué hace el actor** | **Qué hace el sistema** |
|---|---|
| El introduce su login y su password en el sistema | El sistema le devuelve la confirmación de su password. |
| Llena el formulario con los datos que se piden | El sistema devuelve la confirmación de que el perfil o formulario ha sido llenado correctamente. De lo contrario devuelve un mensaje donde informa al usuario que los datos no han sido llenados correctamente y envía este formulario al procesador de la información |
| El usuario realiza una búsqueda de documento | La ontología le devuelve un documento WWW. |
| El usuario decide obtener los elementos esenciales del documento | El sistema le devuelve los elementos esenciales del texto |
| El usuario desea obtener un resumen de las partes esenciales del texto | El sistema le devuelve un resumen. Un documento texto a texto completo y las URL de los documentos |

Tabla. 10. Diagrama de Eventos Usuario del Sistema

El usuario introduce su pasword y su login y el sistema inmediatamente le envía una información de confirmación y este llena el formulario. Si el usuario está registrado solo tiene que teclear su login y password y busca los documentos que necesita resumir, selecciona los segmentos de la estructura textual y obtiene el resumen.

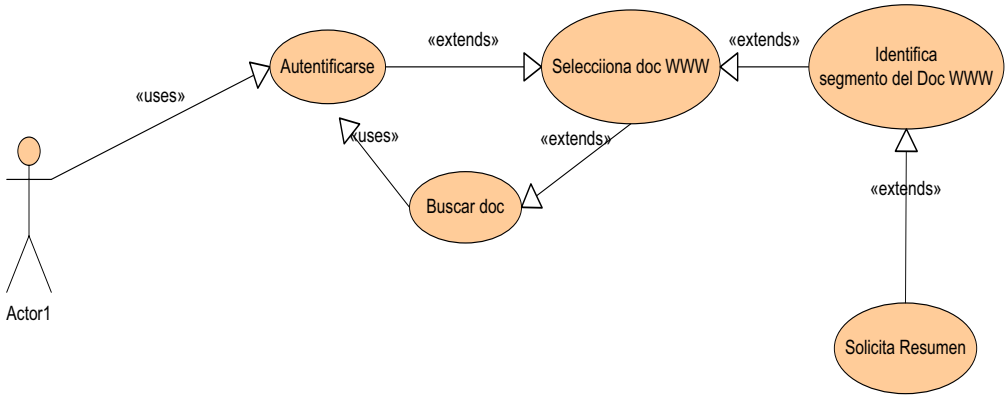



Figura 28. Caso de Uso del Usuario

El procesador inserta su login y su password en la máquina e inmediatamente el sistema le envía una información de confirmación. El sistema le muestra al usuario los datos de los formularios y este entonces comienza a insertar los datos en cada segmento de la ontología, declara las asociaciones y anota los recursos que ha formulado el usuario. Envía los formularios a la ontología ya actualizados. El procesador puede también mediante su login modificar los datos que desee en la ontología (Ver Figura 29 y tabla 11).

| **Qué hace el actor** | **Qué hace el sistema** |
|---|---|
| Inserta su login y password en el sistema | El sistema le devuelve la confirmación de su login y su password |
| Revisa los formularios recibidos | Muestra los formularios recibidos |
| Envía los formularios a la ontología | Recibe los formularios analizados y los integra a la ontología |
| Modifica los elementos que necesiten modificarse o actualizarse para mantener la ontología actualizada. | Devuelve una confirmación de actualización. |

Tabla 11. Diagrama de Evento Rol del Procesador

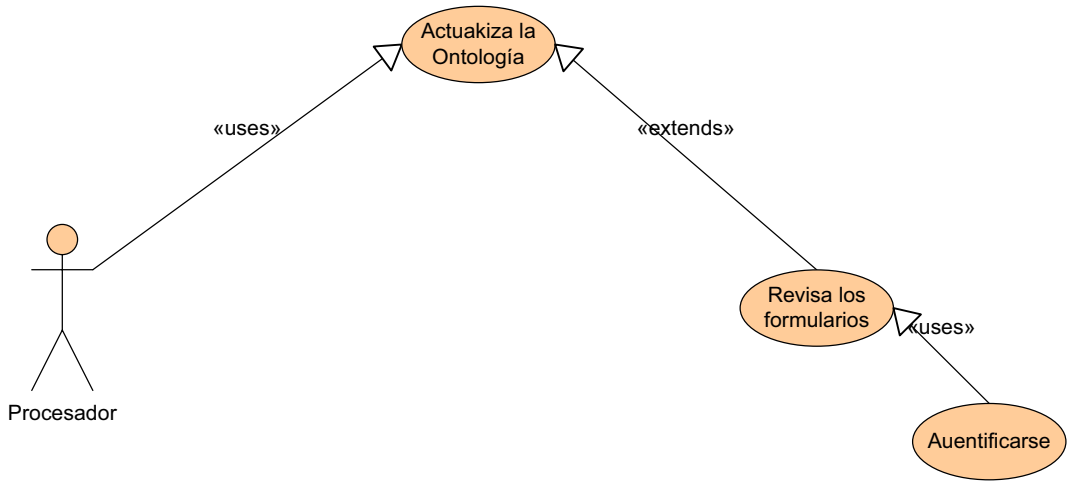



Figura 29. Caso de Uso del Administrador

El administrador del sistema tiene la opción única de otorgar permisos a los usuarios del sistema para que desarrollen su actividad. Se reconoce que estos gráficos no componen de forma íntegra los procesos de ingeniería de software para el sistema, aún faltan diagramas de estado, de clase y de secuencia, los cuales se abordarán en la tesis doctoral, pues en esta investigación solo se aborda un modelo teórico, no un modelo operacional o de sistema (Tabla 12, Figura 28).

| **Que hace el actor** | **Qué hace el sistema** |
|---|---|
| Se autentifica mediante el pasword y login | El sistema devuelve un mensaje de confirmación |
| Otorga permisos para procesar documento | Confirma sobre la concesión del permiso de usuario. |

Tabla 12. Tabla de estados del administrador

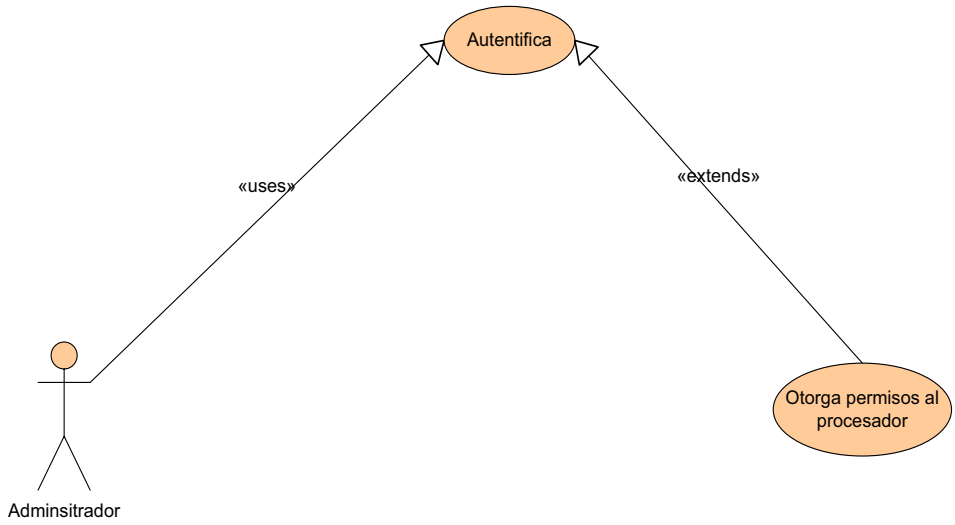



Figura 30. Casos de Uso de Administración de Sistema

### 4.4.9.- Limitaciones

El metamodelo que se ha propuesto tiene limitantes que deben ser salvadas para su implementación, en este casos tenemos como restricciones las siguientes:

- Este procedimiento necesita del concurso de un equipo especializado y multidisciplinario que ejecute el proyecto, pues el mismo tiene elevados niveles de programación.
- Los agentes empleados en el proceso de textos tienden a ser lentos, pues a veces la carga de textos que existe en la ontología es tal que se demoran las operaciones.
- Se necesita un equipamiento especializado para el desarrollo de este sistema, por lo que se hace necesario contar con diversos software y máquinas para el proceso.
- El Metamodelo, necesita aún de un modelo asociado que permita representar la información que se obtiene de la ontología, pues es muy engorroso y poco eficiente determinarlo por los métodos de bolsas de palabras eminentemente estadístico o mediante métodos sintácticos, pero hasta el momento son los más eficaces para el desarrollo del sistema.

### 4.4.10.- Competencias necesarias para el desarrollo del metamodelo

Un metamodelo no puede obviar el componente de los recursos humanos que realizan la representación Hernández (Hernández, 2007). Inobjetablemente las condiciones que exige este modelo de resumen demandan un culmen de habilidades que obliga a los representadores de información a su inserción en el análisis de dominio discursivo mediante estrategias lingüísticas, matemáticas, semiológicas y cognitivas.

Lograr el desarrollo de habilidades en el tratamiento morfológico, lexicográfico, psicológico debe contribuir al desarrollo de conocimientos suficientes para desplegar aún más el campo del resumen automático. Según Hernández (Hernández, 2007) se necesita incrementar los conocimientos teóricos sobre el andamiaje lingüístico para que no solo se describa a través de una redacción técnica, sino para que se interactúe amigable e inteligentemente con los dominios discursivos. Las habilidades para desmontar textos apoyados en competencias gramaticales, semánticas y contextuales, también necesitan del dominio de herramientas de algoritmación, lógica difusa y teoría de conjuntos, que si bien se desarrollan en los primeros años de la academia necesitan de una connotación contextual para ser comprendidas como herramientas de valor en diversos procesos de la BCI (Bibliotecología y Ciencia de la Información).

Para lograr estos objetivos se demanda del conocimiento de disciplinas entre las que se encuentran las siguientes:

- Semántica discursiva, el análisis del discurso.
- Teorías sobre la lectura y la comprensión textual.
- Terminología y terminografía.
- Teoría de la comunicación.
- Lógica Difusa.
- Algoritmación.
- Minería Textual.

El análisis de dominio en estos marcos es una herramienta cuyo nivel intelectual está demandando una perspectiva interdisciplinar, además del

conocimiento de redes sociales. Este modelo exige resumidores identificados con las transformaciones de los modelos matemáticos, semánticos y cognitivos. Según Hernández (Hernández, 2007) consultores atentos a los cambios en las tendencias investigativas, comunicadores imbuidos de la necesidad de una actualización constante en cuanto a técnicas y herramientas, gestores de análisis crítico de los recursos de conocimiento y auditores constantes, modestos e implacables de los sistemas de organización y representación.


# Referencias Bibliográficas

***Referencias Bibliográficas***

AGUIRRE, E. 1998. *Formalization of Concept-Relatedness Using Ontologies: applications in the construction of lexical knowledge bases, word sense disambiguation and automatic spelling correction, .* A dissertation in Computer Science.

ARCO, L. 2008. *Agrupamiento basado en intermediación diferencial.* Tesis Doctoral, Universidad Central"Marta Abreu" de las Villas.

BÉLANGER, A. 2005. *Theory of summarization,* Canadá.

BERRY, M. 2004. *Survey of Text mining: Clustering, Classification, and Retrieval,* New York, USA, Springer Verlag.

BIRKLEY, T. 2000. *Models of ontolgy*, Willey.

BOLÍVAR, A. 1998. *El discurso estilístico en el periodismo.* Tesis de Grado.

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

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DOMÍNGUEZ, S. 2011. FOXCORP. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de Ingeniería Automática.

ENDRES-NIGGEMEYER, B. 2005. SimSum: an empirically founded simulation of summarizing *Information Processing and Management,* 36**,** 659-682.

ENDRES-NIGGEMEYER, B., MAIRE, E. Y SIGEL, A. 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing & Management,* 31**,** 631-674.

FABER, P. & MAIRAL-USÓN, R. 1999. *Constructing a lexicon of English verbs,* Berlín, Mouton de Gruyter.

FABER, P., MONTERO MARTÍNEZ, S., CASTRO PRIETO, M. R., SENSO, J. A., PRIETO VELASCO, J. A., LEÓN ARAUZ, P., MÁRQUEZ LINARES, C. & VEGA EXPÓSITO, M. 2006. Process-oriented terminology

management in the domain of coastal engineering. *Terminology* 2**,** 189-213.

FENSEL, D. 2000. OIL in a nutshell. *In Proceedings of 12th European workshop knowledge acquisition, modelling and management.* New York: Springer-Verlang.

FILLMORE, C. 1982. Frame Semantics. *In:* KOREA, T., L. S. O (ed.) *Linguistics in the morning calm.* Seoul, Hanshin.

FILLMORE, C. & ATKINS, S. 1998. FrameNet and Lexicographic Relevance. *Frist Internacional Conference on Languages Resources and Evaluation.* Granada.

FILLMORE, C., C., J. & PETRUCK, M. 2003. Background to Frame- Net. . *Journal of lexicography,* 16**,** 235-250.

FRAKES, W. & BAEZA-YATES, R. 1992. *Information Retrieval :data Structure & Algorithms,* New York.

GAHL, S. 1998a. Automatic extraction of subcategorization frames for corpus-based dictionary making. *In Proceedings of Euralex'98*

GAHL, S. Year. Automatic extraction of subcategorization frames for corpus-based dictionary making. *In:* In Proceedings of Euralex'98 1998b. 445-452.

GILL, K. & WHEDBEE, S. 2003. *Pragmatic in discurse*, Interamericana.

HALLIDAY, M. & HASAN, R. 1976. *Cohesion in English,* Essex, Longman.

HAYS, D. 1960. *Basic Principles and Technical Variations in Sentence Structure Determination,* Santa Mónica, RAND Corporation (Mathematical Division).

HAYS, D. 1964. *Dependence Theory: A Formalism and Some Observations Language 40*.

HENDLER, A. 2000. Sistems of text extraction.

HERNÁNDEZ, A. 2006. indización y Resumen. La Habana: Universidad de la Habana.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

HJØRLAND, B. 2004. Domain analysis in information science. *Encyclopedia of Library and Information Science.* New York: Marcel Dekker.

IRAZAZÁBAL, A. 1996. Terminología y documentación. *Jornada panllatina de terminologia: perspectives i camps d'aplicación.*

LABOV, W. & WALETSKY, J. 1967. Narrative analysis: Oral versions of personal experience. *Essays on the verbal and visual arts.* Seattle: University of Washington Press.

LANDAUER, T. & DUMAIS, S. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review,* 104**,** 211-240.

LANQUILLON, C. 2002. Enhancing Text Classification to Improve Information Filtering. *Künstliche Intelligenz***,** 37-38.

LASSILA, O. & SWICK, R. 1999. Resource Description Framework (RDF) Model and Syntax Specification. *W3C Recommendation.*

LEWIS, D. D. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *15th Ann Int'1 SIGIR '92.* Denmark.

LÓPEZ-HUERTAS, M. 2008. Organización y representación del conocimiento: curso de doctorado. La Habana: Universidad de la Habana.

LÓPEZ-RODRÍGUEZ, C., TERCEDOR, M. & FABER, P. 2006. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. *Revista eSauld.com,* 2.

LYONS, J. 1977. *Semantics,* London, Cambridge University Press.

MANN, W. & THOMPSON, S. 1990. Rhetorical structure theory: a theory of text organization.

MANNIG, C. & SHÜTZE, H. 2000. *Foundations of Statistical Natural Language Processing,* Cambridge, MA, MIT Press.

MARCU, D. 1997. The Rhetorical Parsing of Natural Language Texts. *In Proceeding of 35th Annual Meeting of the Association for Computational Linguistics, (ACL'97/EACL'97).* Madrid: ACM

MCENERY, A. & WILSON, A. 1996. *Corpus Linguistics,* Edimburgo, Edimburg University Press.

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

MEL'CUK, I. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. *In:* AGEL, V., EICHINNGER, L., EROMS, H., HELLWIG, P., HERRINGER, H. J. & LOBIN, H. (eds.) *Dependency and Valency :an International Handbook of Contemporary Research.* Berlín - Nueva York: W. de Gruyter.

MILLER, G. 1993. *Introduction to WordNet: An On-line Lexical Database.* [Online]. Available: Disponible en: http://elies.rediris.es/elies9/2-4-2.htm [Accessed 26.mayo 2008].

MITKOV, R. 1998. Robust pronoun resolution with limited knowledge. *Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98* Conference.: Montreal.

MLADENIC, D. & GROBELNIK, M. 1998. Feature selection for classification based on text hierarchy. *Conference on Automatic Learning and Discovery (CONALD-98).*

MOREIRO, J. 2004. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural,* Madrid, Ediciones Trea.

MÜLLER, R., SPILIOPOULOU, M. & LENZ, H. 2005. The influence of incentives and culture on knowledge sharing. *In proceeding of 38th Hawaii International Conference on System Sciences (HICSS-38 2005).* Big Island, Hawaii, USA IEEE Computer Society.

NIGAM, K., MCCALLUM, A., THRUN, S. & MITCHELL, T. 2000. Text classification from la belled and unlabeled documents using EM. *Machine Learning,* 38**,** 103-115.

NÚÑEZ, I. 2005. *AMIGA.* Tesis Doctoral, Universidad de la Habana.

ONO, K., SUMITA, K. & MIIKE, S. 1994. Abstract generation based on rhetorical structure extraction. *Proceedings of the International Conference on Computational Linguistics.* Kyoto.

PÉREZ-ÁLVAREZ, J. 1998. *Introducción a la información y documentación científica,* Madrid, Alhambra/Universidad.

PINTO, M. & CORDÓN, J.-M. E. 1999. *Técnicas documentales aplicadas a la traducción,* Madrid, Síntesis.

PORTER, M. 1980. An algorithm for suffix stripping. *Program,* 14**,** 130-137.

RAMÍREZ, Z. 2007. *El Análisis del dominio en la organización y representación del conocimiento.* Diploma de Estudios Avanzados, Universidad de Granada.

RESNIK, P. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships. Doctoral Dissertation.*

RESNIK, P. 1995 , . Using Information Content to Evaluate Semantic Similarity in a Taxonomy". *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI).*

RIGAU, G. 2002. Resolución automática de la ambigüedad semántica de palabras. , Fundación Duques de Soria, curso de Tecnologías de la lengua.

RIGAU, G. 1998, "Automatic Acquisition of Lexical Knowledge from MRDs", PhD Thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona.

ROBINSON, J. 1970. A Dependency Based Transformational Grammar. *Actes du Xme Congrès international des linguistes 2.* Bucarets.

SAHAMI, M., DUMAIS, D., HECKERMAN, D. & HOVITZ, E. 1988. A Bayessian approach to filtering junk a-mail.

SALES-SALVADOR, D. 2006. *Documentación aplicada a la traducción: presente y futuro de una disciplina,* Gijón, Trea.

SALTON, G. & BUCKLEY, C. 1988. Term weighting approaches. *Automatic text Information Processing and Management,* 24**,** 513-523.

SALTON, G., WONG, A. & YANG, C. S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,,* 18**,** 613-620.

SANDIG, L. & SELTING, B. 2003. *Los retos del discurso espistémico,* México, Interamericana.

SCHAAL, M., MÜLLER, R., BRUNZEL, M. & SPILIOPOULOU, M. 2005. RELFIN - Topic discovery for ontology enhancement and annnotation. *In:* GÓMEZ-PÉREZ, A. & EUZENAT, J. (eds.) *In Proceeding of The Semantic Web: Research and Applications, Second European Semantic Web Symposium (ESWC 2005).* Crete, Greece: Springer, Heraklion.

SEARLE, J. 1965. What is a speech act *In:* BLACK, M. (ed.) *Philosophy in Americ.* Londres: George Allen.

SEARLE, J. 1969. *Speech acts. An essay in the Philosophy of Language,* Cambridge, Cambridge University Press.

SEARLE, J. 1975. Indirect speech acts. *In:* COLE, P. & MORGAN, J. (eds.) *Syntax and Semantics 3. Speech Acts.* New York: Academic Press.

SENSO, J. A., MAGAÑA, P. J., FABER-BENITEZ, P. & VILA, M. M. 2007. Metodología para la estructuración del conocimiento de una disciplina: el caso de PuertoTerm. *El profesional de la Información,* 16**,** 591-604.

TESNIÈRE, L. 1959. *Éléments de syntaxe structurale,* París, Klincksieck.

TOMLIN, K. 2003. *Pragmatic and semantic in discurse*.

VAN DIJK, T. 1978. *La Noticia como discurso: comprensión, estructura y producción de la información,* Barcelona, Paidós.

VAN DIJK, T. 1980. *Estructura y funciones del discurso,* México, D.F., Siglo veintiuno.

VAN DIJK, T. 1984. *Texto y Contexto: Semántica y Pragmática del discurso,* Madrid, Cátedra.

VAN DIJK, T. 1995. De la gramática del texto al análisis crítico del discurso. *BELIAR,* 2.

VAN DIJK, T. 2004. *Discurso y desigualdad,* Tenerife, Universidad de La Laguna.

VERONIS, J. 2000. Sense tagging: Don't look for the meaning but for the use", Computational Lexicography and Multimedia Dictionaries (COMLEX'2000), pp. 1-9.

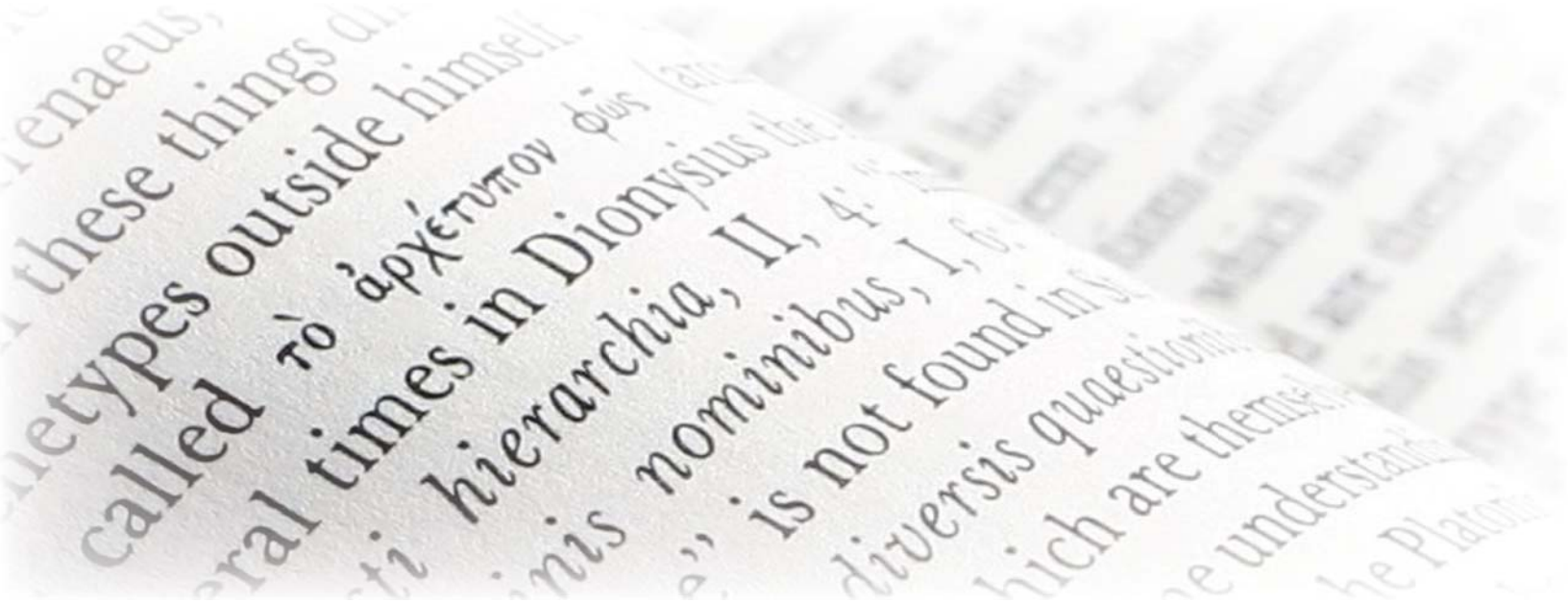
VILLENA, J., GONZALEZ, J. & GONZALEZ, B. 2002. Dedalus: modelo de desambiguación.

YANG, Y. & PEDERSEN, J. 1997. A comparative study on feature selection in text categorization. *Journal of Artificial Intelligence Research,* 6**,** 1-34.

YAROWSKY, D. 2000. Hierarchical Decision Lists for Word Sense Disambiguation, *Computers and the Humanities*. Special Issue: Evaluating Word Sense Disambiguation Programs, 34 (1-2), 179-186.

ZALDUA, A. 2006. El análisis del discurso en la organización y representación de la información-conocimiento: elementos teóricos. *ACIMED* 14.

ZIDOROV, G. & OLIVA, O. n.d. *RE: Resolución de anáfora pronominal para el español usando el método de conocimiento limitado*.*




# Capítulo V

## Análisis del Discurso Científico en dominio de la Ingeniería de Puertos y Costas: sus implicaciones para la construcción del software

# Capítulo 5: Análisis del Discurso Científico en el dominio de la Ingeniería de Puertos y Costas: sus implicaciones para la construcción del software.

**Introducción:**

En este apartado de la investigación se describe el análisis de los textos del dominio Ingeniería de Puertos y Costas con el objetivo de detectar regularidades lingüísticas, discursivas y sintáctico-comunicativas para ser implementadas en el sistema de resúmenes, para ello se han seleccionado los corpus del proyecto PuertoTerm de la Universidad de Granada, liderado por la Dra. Pamela Faber Benítez (Faber et al., 2005).

Se reconoce que el análisis de los corpus tiene una connotación especial en el tratamiento del texto. En él están desarrollados y explicitados elementos esenciales que facilitarán el reconocimiento de entidades lingüísticas, de elementos macroestructurales, microestructurales y macroreglas por los ordenadores, así como el análisis de la retórica y el estilo.

Finalmente en este capítulo se muestran los elementos necesarios para la implementación del modelo sobre textos científicos a partir de los procederes metodológicos aplicados y desarrollados por Iria D´Cunha Fanego (D'Cunha, 2006), Brigitte Enders Niggemeyer (Endres-Niggemeyer, 2005, Endres-Niggemeyer, 1995) y Mann y Thopmson (Mann and Thompson, 1990) de los cuales se utilizan las herramientas de análisis esenciales de este capítulo.

## *5.1- Criterios de selección del Corpus*

Como declara Senso (Senso, 2009b), para el análisis de corpus se necesitan técnicas de recuperación de información asociadas a la lingüística de corpus, parcela disciplinar cuyo valor instrumental es palpable en proyectos a gran escala como las dos fases de ACQUILEX o, a menor escala dentro del mundo de los lenguajes documentas, los que pretenden generar clasificaciones, tesauros u ontologías (ISO-5964, 1985).

La definición de corpus centra sus bases en las concepciones teóricas de Bowker y Pearson (Bowker and Pearson, 2002), quiénes aseveran que éste puede ser descrito como *una gran colección de textos auténticos que han sido compilados de forma electrónica de acuerdo a una serie de criterios previos*(Bowker and Pearson, 2002). Otras aristas de este concepto están sustentadas por la lingüística en la cual se define a los estudios de corpus como *el estudio del lenguaje basado en ejemplos extraídos a partir del uso del idioma en la "vida real"*(Mc Enery and Wilson, 1996). Según Senso (Senso, 2009a) Biber (Biber et al., 1988) y sus colaboradores refieren que los estudios de corpus consisten en análisis empíricos de una amplia colección de textos reales, para los que el investigador hace uso extensivo de material informático y aplica técnicas de análisis tanto cuantitativas como cualitativas. Con todas estas características, la lingüística de corpus atraviesa, hoy por hoy, un momento de creciente popularidad dentro del ámbito de estudio de la lexicografía (Biber et al., 1998), no así en el terreno de la construcción de extractos, donde la impronta de la Ciencia de la Computación ha penetrado en los procesos de construcción de grupos de textos desentendiéndose en muchos casos de los postulados de la lingüística para dar preponderancia a los algoritmos de clustering.

### 5.1.1-. Tipologías de Corpus

Los corpus tienen una alta connotación en el desarrollo de sistemas de extracción de texto. Existe una gama, cada vez más variada, de corpus (Bowker and Pearson, 2002):

1. Corpus genéricos frente a corpus especializados
2. Corpus escritos frente a orales
3. Corpus monolingües frente a multilingües
4. Corpus sincrónicos frente a diacrónicos
5. Corpus abiertos frente a cerrados
6. Corpus de aprendices
7. Corpus impresos frente a los electrónicos

Estas tipologías de corpus facilitan la formulación de disímiles análisis en diversos contextos como el discursivo y el terminológico. Los corpus han servido como medios de comprobación de la existencia de subdisciplinas asociadas a la Minería Textual, la Representación de la Información y el Conocimiento. Se trata de la gramática, la semántica, la pragmática, el análisis discursivo, la lexicografía, la sociolingüística, la lingüística histórica, la traducción y la documentación (Senso, 2009a).

Estas parcelas de conocimiento defienden la posición del análisis de corpus como una unidad netamente descriptiva (frente a las aproximaciones de abstracción propias de la gramática generativa) que se empeña en identificar las asociaciones que rigen el funcionamiento de los diversos componentes textuales:

1 Análisis de Macroreglas: Declaración de elisiones y generalizaciones.

2 Análisis de Microreglas: Deducción de todos los elementos que se integran a la estructura sintáctica del texto tanto a nivel oracional como de párrafo.

3 Análisis de Contexto: permite determinar la realidad comunicativa en que se encuentra el texto y el usuario.

4 Análisis de Estilo y Retórica: Permiten obtener los elementos que caracterizan el discurso del autor en y la identificación de aquellas cuestiones que se utilizan para hacer creíble el artículo científico.

5 Análisis Macroestructural: Facilita la identificación de los tópicos del texto y los ejes semánticos hacia donde se enruta el texto.

6 Análisis Microestructural: Permite obtener la primeras palabras y diversas categorías gramaticales que conforman el texto.

Estas técnicas son muy útiles, pues son posturas de análisis de discurso que suponen una vía para reflejar todos los elementos discursivos asociados al corpus, los ejes temáticos, los vocablos y sus niveles sintácticos y sintagmáticos. Como enuncia (Senso, 2009a), desde el punto de vista de uso, la utilización real de los vocablos permite que el investigador se asegure que la

información extraída es fiel reflejo de los contenidos reales. El caso que nos ocupa es el reflejo de documentos asociados al dominio de la Ingeniería de Costas, y del sublenguaje especializado empleado en el mismo.

### *5.2. Características del proyecto Proyecto Puerto Term*

La Ingeniería de Puertos y Costas es una disciplina que difiere del resto de las ingenierías. La inexistencia de diccionarios específicos, la ausencia de teorías específicas para este campo multidisciplinar y la necesidad que tienen los profesionales que en ella trabajan de poseer herramientas terminológicas y léxico semánticas para realizar su trabajo le confiere un cierto halo de área discriminada al menos, hablando desde el punto de vista de la documentación y la terminología (Leiva et al., 2009a, Senso, 2009a, Senso and Leiva, 2008, Leiva et al., 2009b).

Los objetivos específicos sobre los que sustentó el proyecto fueron según Senso (2007):

1 Establecer un corpus de textos concretos para esta disciplina en español, inglés y alemán.

2 Especificar cuales serían los conceptos y los términos que desarrollan la arquitectura semántica de esa disciplina y establecer las relaciones conceptuales específicas.

3 Diseñar y alimentar una base de conocimiento terminológico articulado en torno a la estructura hallada en la definición de los términos. En ella, además, se deberían de poder almacenar las relaciones semánticas existentes entre los términos, de forma que posteriormente se pueda producir una conversión o vinculación a formas de representación más expresivas.

4 Crear una aplicación informática que permitiera la recuperación de la información sin que se perdiera la estructura y relaciones formadas entre los conceptos y los términos.

5 Crear un banco de imágenes para complementar y enriquecer las representaciones lingüísticas de los conceptos pertenecientes al campo

especializado que, por sus características específicas, necesita una representación conceptual más visual y dinámica.

A estos objetivos se suma otro empeño del dominio discursivo relacionado con este tema: la necesidad de desarrollar un sistema de información que englobase todos los elementos anteriormente referidos y ofreciera resúmenes extractos-abstractos de textos utilizados en el corpus.

### 5.2.1- Representatividad del corpus

No se conocen los criterios exactos para seleccionar un corpus ni para su diseño. Según Sánchez (Sánchez et al., 1995) para que sea un corpus realmente representativo del dominio discursivo que representa es importante preguntarse lo siguiente: ¿Qué complejidades de uso del habla debe incluir? ¿En qué compensación? ¿Cuál debe ser la dimensión de un corpus para que, realmente, represente un dominio lingüístico? ¿Qué nivel de expresividad posee ese corpus?, ¿Qué cualidades léxico-semánticas tiene el corpus?

Estas formulaciones son las que guían los criterios de recopilación de los textos incluidos en el corpus de esta investigación, así como el empleo de técnicas matemáticas que permitan determinar la calidad del corpus para su selección (Capítulo 7). Como apuntan Senso, Biber, McEnery y Wilson (Senso, 2009a, Biber et al., 1998, Mc Enery and Wilson, 1996), aunque la literatura sobre este campo es extensa, la realidad hasta el día de hoy, es que, casi todos los córpora que se han diseñado con criterios que favorezcan las aplicaciones han hecho que las mismas sean muy eficientes. Solo en algunas investigaciones relativas al campo de la lingüística del corpus como el (British National Corpus, Birmingham Collection of English Text, Corpus CUMBRE o el Corpus ARTHUS) se han explicitado los criterios de selección de los textos incluidos en el corpus (Sánchez et al., 1995).

Es indiscutible que elementos como: representatividad, estandarización y la tipología de los corpus han sido temas de constante análisis dentro del gremio científico. La mayoría de las enunciaciones emanan del seno de significativos proyectos europeos (Atkins, 1992, Senso, 2009a). Según Senso (Senso, 2009a) en EAGLES, por ejemplo, Sinclair define un conjunto de elementos

indispensables, que deben asumirse como norma para textos en formato electrónico (cantidad, calidad, simplicidad y documentación), y clasifica los diferentes tipos de córpora que pueden existir, para así diferenciarlos de las colecciones de textos o los archivos (archives), ya que estos últimos no cumplen alguna de ellas:

1 El corpus debe ser tan grande con alto nivel de representatividad, en esta época las condiciones tecnológicas permiten conjuntos textuales de considerables dimensiones.

2 Heterogeneidad, debe contener ejemplos de gran cantidad de géneros textuales o discursivos, con el fin de contemplar todas las posibles representaciones.

3 Debe ser una clasificación intermedia entre géneros.

4 Los ejemplos (cada fichero) debe tener un tamaño representativo.

5 Deben tener una descripción a nivel de metadatos clara, válida y actual.

A pesar de estos criterios, la mayoría de los corpus desarrollados para ser usados en sistemas lingüísticos que utiliza la comunidad científica no cumplen con estas directrices, aunque se evidencia que algunos proyectos comienzan a seguir los lineamientos declarados por EAGLES. La discusión procedimental está sentada sobre el uso de dos indicadores: calidad vs. cantidad sigue siendo el caballo de batalla a la hora de desarrollar corpus textuales. Por un lado están aquellos que dan más importancia al hecho de que el corpus fuera representativo y equilibrado y por otro aparece el criterio de aquellos que, además, destacan que el valor de un corpus está esencialmente en que el corpus fuese lo más cuantioso posible (Péres Hernández, 2002).

El autor sostiene su criterio de selección teniendo en cuenta dos cuestiones: la regularidad que dicta que cuanto mayor sea el corpus, más posibilidades semánticas y lexicográficas mostrará y las evaluaciones de la calidad del corpus mediante técnicas matemáticas y de agrupamiento declaradas en el Capítulo 7 (Ver Capítulo 7). Según Senso y Faber (Senso, 2009a, Faber et al., 2005), en su análisis de los corpus lexicográficos –corpus para lexicografía-

"esto último puede ser determinante en algunos casos, por ejemplo, en la lexicografía: un diccionario como el OED contiene 250.000 entradas y un diccionario medio para estudiantes de 50.000 palabras, por lo que acumular evidencias lingüísticas (al menos las suficientes como para poder guiar al lexicógrafo en el proceso de compilación) sobre un número tan elevado de entradas requiere, sin duda, que el corpus sea, por decirlo de alguna forma, cuanto más grande mejor. Si pensamos por ejemplo en el estudio de los hábitos colocacionales de determinadas palabras, cuanto mayor sea el volumen de texto que procesemos, más representativos serán los índices estadísticos de frecuencia que aparezcan, al ser relativos a una mayor cantidad de texto.

El crecimiento de la información en determinadas áreas dentro de la web ha generado un aumento desigual en lo que se refiere a la aparición de herramientas terminológicas que faciliten la recuperación de información (taxonomías, tesauros y ontologías). Así, dependiendo del nivel de asentamiento que tenga una disciplina determinada, será más o menos complejo localizar productos que permitan estructurar el conocimiento asociado a dicha rama del saber.

Un ejemplo claro lo encontramos en la Ingeniería de Puertos y Costas. Aunque se trata de una disciplina relativamente nueva, mezcla de la clásica Ingeniería de Caminos, Canales y Puertos y las ingenierías oceánica y naval, la notable evolución científica que ha experimentado debido al desarrollo de nuevas técnicas de construcción, explotación y gestión de estructuras y recursos marinos no ha sido suficiente como para que exista una formalización de las organizaciones de conocimiento que representen los procedimientos, definiciones, objetos y el metaconocimiento que en ella se llevan a cabo.

Esto es fácilmente comprobable si se realiza una consulta simple sobre dos tipos de herramientas de ontologías, buscando por los términos más representativos relacionados con cuatro disciplinas escogidas de forma aleatoria dentro de la Ingeniería (Agrícola, Geológica, Aeronáutica y Civil) a la que sumamos la Ingeniería de Puertos y Costas. El resultado será, por lo tanto, el número total de clases dentro de las ontologías que tratan, de una manera u

otra, aspectos relacionados con dichas disciplinas y, por lo tanto, que permitan representar el conocimiento inherente a ellas. Las herramientas empleadas para realizar las consultas han sido:

1 Buscadores de ficheros OWL y, en general, cualquier tipo de lenguaje para ontologías:

- Swoogle
- Falcons
- Intellidimension
- Dr Watson

2 Repositorios de ontologías:

- OwlSight Ontology Repository
- OpenOntology Repository
- Tones

Y los resultados, plasmados en el siguiente gráfico, demuestran la poca formalización que tiene la disciplina de Ingeniería de Puertos y Costas, al menos en "formato ontología", en relación con el resto dentro de su misma área de actuación (Ver figura 31). Como es bien sabido, no existen herramientas que permitan realizar este tipo de búsquedas sobre tesauros.

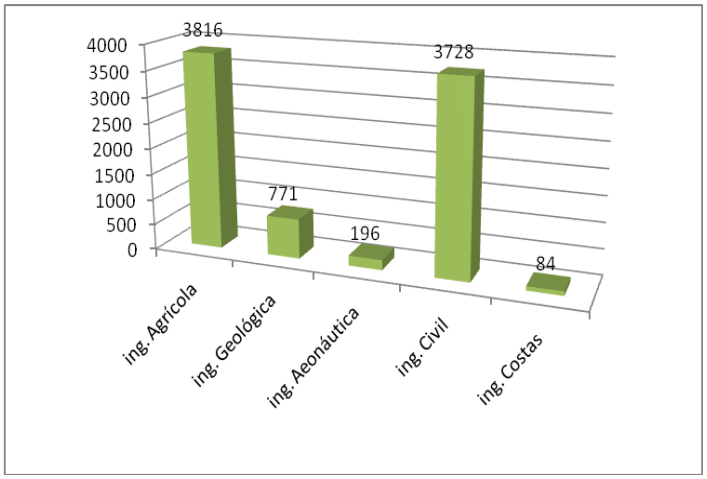



Figura 31. Temáticas esenciales en las bases de datos

La mayoría de miembros de dicho proyecto de investigación ha formado parte de otros dos proyectos anteriores, dentro de la misma temática, que han servido como base teórica y, especialmente, metodológica, para el que se explica a continuación. El primer proyecto sirvió para desarrollar una lógica léxica para la traducción asistida por ordenador a partir de una base de datos y el segundo, de gran impacto dentro de la comunidad médica, permitió la elaboración de un sistema de información integrado en internet y dedicado al subdominio de la oncología. Su nombre fue Oncoterm.

En la actualidad, el corpus con el que se está trabajando en el proyecto PuertoTerm, y que se empleará también en esta tesis, contiene más de 10.000 ficheros en total, siendo los del español (3.535) y el inglés (4.861) los más abundantes sobre los de alemán (1.786). Para desarrollar el experimento de esta investigación se seleccionó el corpus español y a continuación se declaran las razones de su elección:

1. Ausencia de corpus en español debidamente estructurado para la confección de extractos en el terreno de la ingeniería de puertos y costas.
2. Existencia de un grupo importante de investigadores que hablan español sobre todo en América Latina.
3. Posibilidad de desarrollar una aplicación cuyas dimensiones puedan aprovecharse en proyectos en lengua hispana.
4. Existencia de información valiosa en español sobre el tema que se declara, la cual no puede ser estructurada ni difundida por falta de tratamiento.

Para desarrollar la aplicación que se propone se seleccionó un subcorpus en español contentivo de 50 artículos de forma intencional de manera que todos tuviesen en su estructura introducción, desarrollo, metodología, resultados y conclusiones, dada la necesidad de extraer información de fuentes en este idioma y la escasez de las recurso de información en el terreno de la Ingeniería de Puertos y Costas.

### 5.2.1- Características del Corpus de PuertoTerm

Los corpus que han servido sustento para esta investigación poseen características muy disímiles en lo referente a nivel discursivo, prestigio de sus fuentes y tipología documental. Esta situación hace que las estrategias de trabajo y análisis sean más complejas, debido a la disparidad en la concepción del corpus.

Si se analiza el corpus desde el punto de vista de estructura discursiva puede llegarse a la conclusión de que el corpus inglés es superior, pues el discurso que genera es eminentemente especializado (más de 4000) registros están en este orden, (818 son semi-especializados) y solo 43 son artículos divulgativos. El corpus español es, en esencia, divulgativo y en menor grado especializado. (Ver figuras 32- 33).

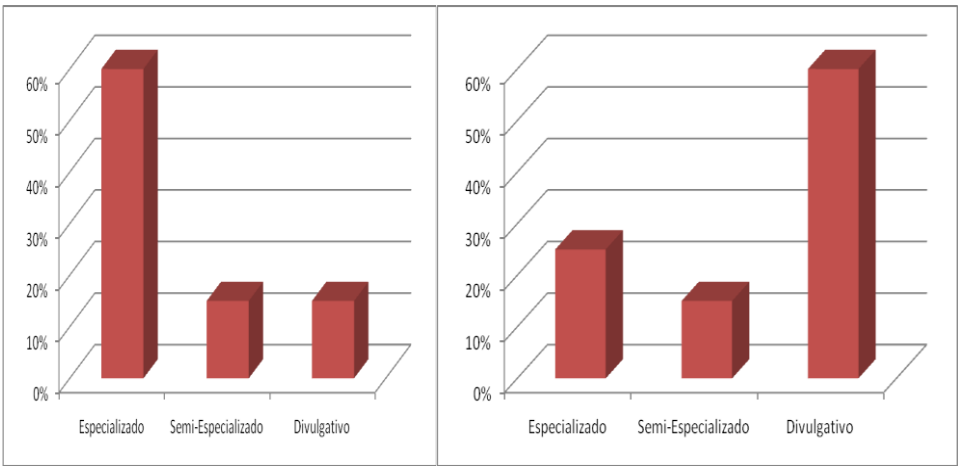



Figura 32: Cualidades del Corpus Inglés y Español

Desde el punto de vista de la procedencia de las fuentes de información son esencialmente de países de habla hispana en el caso del corpus en español (figura 33). Los países que son el sustento del corpus son: España, Venezuela, México, Cuba, Colombia, Costa Rica, Argentina. También hay una cantidad menor de artículos en español generados en los Estados Unidos, Canadá, Holanda, trabajos que adolecen en su capacidad de comunicación, son traducciones de baja calidad con respecto al español.

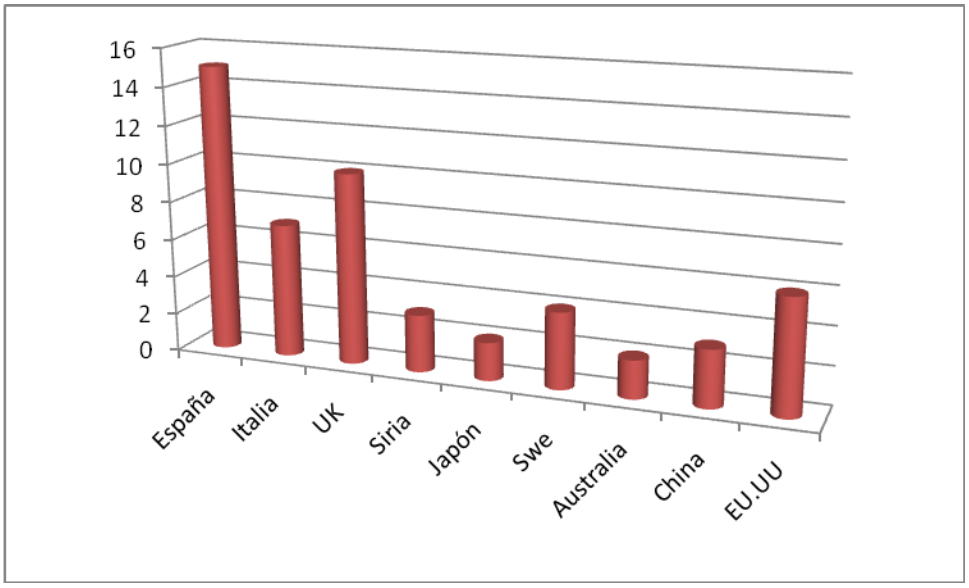



Figura. 33 Países de procedencia del corpus

Las páginas Web que sustentan los elementos que componen el corpus en ingles son publicaciones y bases de datos de prestigio internacional, reconocidas por la comunidad internacional como fuentes de valor para el estudio de la Ingeniería de Puertos y Costas (Ver Tabla 13).

| Listado de las Publicaciones que tributan al corpus en Inglés |
|---|
| Sedimentary Geology 114 (1997) 267-294:The Upper Permian Boniches Conglomerates Formation: evolution from alluvial fan to fluvial system environments and accompanying tectonic and climatic controls in the southeast Iberian Ranges, central Spain |
| Geomorphology 22 (1998) 265-283:Geomorphological and sedimentological analysis of a catastrophic flash flood in the Arfis drainage basin (Central Pyrenees, Spain) |
| Geomorphology 39 2001 3–19: Fluvial geomorphology and paleohydrology in Japan |
| Journal of Hydrology 226 (1999) 48–65:Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework |
| Palaeogeography, Palaeoclimatology, Palaeoecology 151 (1999) 5–37:Late Neogene lacustrine record and palaeogeography in the Quillagua–Llamara basin, Central Andean fore-arc (northern Chile) |
| Tectonophysics 284 (1998) 151-160:Translocated Plio-Pleistocene drainage systems along the Arava fault of the Dead Sea transform |
| Phys. Chem. Earth, Vol. 22, No. 3--4, pp. 345-349, 1997:The Role of Human Activities in the Development of Alluvial Fans |
| Catena 55 (2004) 125–140:Late Quaternary rapid talus dissection and debris flow deposition on an alluvial fan in Syria |

Tabla 13: Fuentes que sirven de base al corpus inglés

Los sitios Web del corpus en inglés pertenecen a instituciones de nivel internacional, mientras que el corpus español está constituido en su mayoría por publicaciones de instituciones de poco prestigio, por lo que los mismos son textos divulgativos, pertenecientes a sitios web de instituciones no científicas, estatales, solo poseen información de recursos de información radicados en España, tal es el caso de TESEO, Aguas, de México, etc. Esta disparidad en la selección de fuentes es perjudicial para el desarrollo de una estrategia de análisis de corpus (Ver tabla 13 y figura 33).

| Fuentes esenciales del Corpus español |
|---|
| Boletín. Instituto Español de Oceanografía |
| Cuaternario y geomorfología |
| Agua |
| TESEO |

Tabla 14 Fuentes principales del corpus español

Para concluir la caracterización del corpus se realizó una aproximación a la estructura de los artículos buscando una paridad en el tratamiento de los textos y en su presentación, para ello analizamos algunas características esenciales de los corpus como: presencia de resumen del autor, estructura científica, claridad en la redacción técnica. Los resultados de este análisis evidencian lo siguiente (Anexo 5).

1 En el corpus español, existe gran cantidad de textos sin resumen, debido a que no son artículos de revista si no materiales divulgativos, no siendo así en el inglés donde abunda el apartado resumen dentro de los artículos.

2 Lo mismo sucede con la estructura la mayoría de los textos en español, es anárquica, pues prácticamente no existe orden, algo que en el corpus es un indicador de excelencia.

3 El otro aspecto que se estudió fue la redacción, si bien este aspecto no es problémico los trabajos del corpus español deben prescindir de localismos y frases poco científicas para llegar más a la comunidad en general.

Se reconoce que el proyecto del que se toma la información no tenía el fin de recoger información para redacción, si no para extraer vocablos y su connotación, sin embargo hubiere sido mejor el resultado si estas cuestiones se hubiesen tenido en cuenta.

A pesar de las dificultades que posee hoy este corpus español, existe la necesidad de usar información en este idioma por gran parte de la población de Hispanoamérica, pues es un dominio muy poco estudiado que posee muy pocos sistemas de información con nivel. Ello ha motivado a que se haya decidido tomar esta fuente para realizar el análisis de corpus, independientemente de las dificultades que pueda acarrear en el tratamiento lingüístico. Las regularidades declaradas en este análisis también pueden extrapolarse al corpus inglés para este tipo de tema.

### *5.3.- Análisis textual: herramientas*

Para realizar dicho proceso, en la lingüística de corpus es común emplear determinadas herramientas informáticas que ofrecen la posibilidad de trabajar con gran cantidad de corpus y procesarlos para generar determinados resultados. Entre ellos, los más empleados son:

1 **Textquest**. Es una herramienta enfocada al análisis de contenido. Admite efectuar exploraciones a partir de patrones de consulta sustentados en palabras, segmentos de vocablos, secuencias de vocablos, y las llamadas cadena de raíces de palabras. El programa funciona etiquetando categorías, lo que implica la descripción detallada de todas las categorías utilizadas. http://www.textquest.de/ .

2 **Concordance**. Sus valores como herramienta de análisis radica en la concordancia lógico-semántica entre archivos o ficheros de texto. Trabaja con diferentes lenguas, de las que elabora listas de palabras y concordancias. También puede trabajar con documentos alojados en páginas web. No realiza análisis estadístico de frecuencias. http://www.concordancesoftware.co.uk/.

3 **TextAnalyst**. Contiene diversas funciones de análisis textual, lo que le facilita la creación de redes semánticas sobre el contenido del corpus textual. Devuelve en sus salidas párrafos textuales con nexos a las estructuras nodales declaradas en la red semántica. Facilita determinar qué conceptos –palabras o combinaciones de palabras- son las más importantes en el contexto del texto bajo estudio. Cada concepto es etiquetado o marcado como un nodo, y se le asigna un “peso numérico” (equivalente a la probabilidad de dicho concepto con relación al texto). TextAnalyst fija los pesos ponderados de los nexos entre conceptos propios en el texto. Esta herramienta tiene la capacidad de leer diversos tipos de formatos como: HTML, Word, txt, etc. http://www.mega-puter.com/textanalyst.php

4 **T-Lab**. Es una herramienta de análisis de texto desarrollada en Italia. Utiliza técnicas de análisis estadístico de texto, minería de texto y análisis

multivariado (análisis de correspondencias y análisis de clúster, entre otros). Permite la extracción, comparación y el mapeo de los contenidos de diversos tipos de textos: transcripciones de discursos, libros, artículos, notas periodísticas, documentos de internet, respuestas a cuestionarios de preguntas abiertas, etc. http://www.tlab.it/es/presentation.php

5 **WordSmith Tools**. Es una aplicación integrada por cuatro grupos de instrumentos principales, desde las cuales pueden realizarse una gran variedad de análisis léxicos utilizando algoritmos y análisis textuales, entre los cuales destacamos:

- **WordList** Facilita la obtención de inventarios de términos ordenadas alfabéticamente o por frecuencia, junto con estadísticas detalladas sobre la composición del corpus. Las listas de palabras pueden estar establecidas sobre unidades léxicas simples o en grupos de dos o más palabras como: verbos sustantivados, adjetivos y sustantivos aislados. Con este software es posible comparar listas de palabras para estudiar diferencias en la frecuencia de su uso e indizar el corpus, con lo que se acelera el tiempo de proceso en las búsquedas y dicho índice puede después usarse para obtener otros tipos de información, como el MI-Score.

- **KeyWords** es un software que permite extraer las "palabras clave" de un corpus, tomando éste como una unidad, o bien de los ficheros que lo componen tomados de forma independiente. Permite, además, estudiar la forma en la que dichas palabras clave se distribuyen el texto y los enlaces que existen entre una palabra clave y las demás.

- **Concord** es un instrumento de concordancia de alto nivel que, además de las particulares especificas de los sistemas de concordancia, admite automatizar colocaciones, análisis de patrones léxicos (patterns) y agrupaciones de palabras (clúster).

- **Viewer**, **Splitter** y **Text Converter** son otras tres herramientas adicionales que permiten, respectivamente, acceder al fichero de texto al que pertenece una línea de concordancia, dividir textos en partes más

pequeñas y convertir el formato de un texto para adaptarlo a las necesidades del usuario.

### *5.4.- Análisis de Discurso*

#### 5.4.1.- Análisis Semántico

Del análisis del discurso ya se habló en el capítulo inicial, esta técnica no constituye un método de investigación particular, si no una herramienta analítica que engloba múltiples métodos de investigación. En este apartado se detallan los ejes semánticos de 50 textos, los que proveerán al sistema de las claves de organización que podrán ser utilizados por la ontología, ya que los ejes semánticos son los elementos que representan las unidades de significado en que se divide el texto y son por tanto una clave de estructuración y clasificación de contenidos (Figura 34).

Para realizar este análisis se han desarrollado los siguientes pasos:

1. Lectura del documento.
2. Seccionar el texto en sus unidades retóricas (Objetivos, Metodologías, Discusión y Resultados).
3. Establecer los unidades temáticas generales y las adyacentes (Detección de unidades núcleo y unidades relacionadas) (Anexo 6).

**Ejemplo:**

| Título y Autores |
| --- |
| Título: *Influencia de la Hidráulica y del viento en la eficiencia de Remoción de Metales Pesados (Cu, Zn Cr y Fe) en una Laguna Facultativa*<br><br>Autores: Aldana Gerardo, Aiello Cateryne, Morán Milmero, Jérez Oswaldo Centro de Investigación del Agua. Universidad del Zulia Apartado 15.380 Delicias. Maracaibo, Venezuela<br><br>Resumen<br><br>*La remoción de metales pesados en tratamiento de agua residual es de gran interés cuando se requiere reutilizar las aguas. Existe una gran variedad de metales pesados en la naturaleza, para efecto de este estudio fueron seleccionados cuatro: Cobre, Zinc, Cromo y Hierro, desde el* |

*punto de vista toxicológico de las aguas. Los objetivos de este trabajo fueron determinar el comportamiento de los metales a diferente profundidad y la eficiencia de remoción para una laguna facultativa, considerando la acción del viento, temperatura, tipo de flujo y la fotosíntesis. La investigación fue ejecutada en la Serie A de las Lagunas de Estabilización de LUZ.*

*Del estudio se observó que las concentraciones de Cu y Fe en las superficies (0,40 m de profundidad) son mayores que las observadas para las muestras tomadas a 2 mts. de profundidad. En el caso del Zinc se observó el mismo comportamiento para el muestreo efectuado a las 7:00 am, mientras que para el muestreo de las 7:00 pm sucede lo contrario. Este efecto puede atribuirse a la acción del viento, el cual ocasiona turbulencia y cambio de velocidad del flujo que hacen que el metal se concentre en la superficie. El porcentaje de remoción obtenido en la laguna facultativa fue de 100, 90, 75 y 64% para el Cobre, Zinc, Cromo y Hierro, respectivamente. El método utilizado fue el de espectrofotómetro de absorción atómica con llama. Palabras claves: remoción, laguna de estabilización, metales pesados, hidráulica*

**Introducción**

La mayoría de los metales pesados con una densidad superior a los 5 g/ml presentes en los organismos vivos, pueden ejercer una acción tóxica en el agua de características especiales según el elemento y la concentración.

Esta toxicidad depende de la capacidad del metal para combinarse y de la unión firme de él con los componentes químicos esenciales de la molécula, desplazando a los metales livianos predominantes en los sistemas vivientes. El propósito del presente trabajo consiste en evaluar el contenido metálico (Cu, Zn, Cr y Fe) a diferentes profundidades en varios sitios de una laguna facultativa. El sistema de lagunas de la Universidad del Zulia, Venezuela se encuentra conformado por tres (3) facultativas y seis (6) de Maduración, las cuales están distribuidas en tres sistemas paralelos de tres lagunas cada sistema (facultativa - 2 maduración). Las dimensiones de cada laguna se indican en la tabla 1. En el análisis de las muestras se empleó la técnica de espectrofotometría de absorción atómica, para un total de 112 muestras durante tres (3) meses.

**Metodología**

Las muestras fueron tomadas en catorce puntos de la laguna A1 facultativa Fig. 1. Dos de los puntos fueron ubicados: uno en la entrada en el interior de la canaleta Parshall y el otro a la salida de la laguna o en el vertedero circular. El resto de los puntos fueron tomados equidistantes uno del otro y de igual forma de los bordes de los taludes; seis de los puntos fueron tomados en la superficie (0,40 m debajo del nivel del agua) y otra cantidad igual a 2 m

de profundidad.

Las muestras fueron tomadas cada 8 días, rotando el día y durante doce semanas. Los horarios seleccionados para la toma de muestra fueron dos: 7:00 am y 7:00 pm. El equipo utilizado para determinar la concentración de los metales fue un espectrofotómetro de absorción atómica Varian Spector 10/20. Los límites de detección fueron 0.003, 0.002, 0.004 y 0.005 ug/ml para el Cu, Zn, Cr y Fe. Las muestras fueron analizadas por duplicado, preconcentrada y comparada con un blanco, su preparación fue según la técnica descrita en el standard método (1989).

**Resultados y Discusión**

En las figuras 1 y 2 se pueden observar los valores de la concentración promedio de las 7:00 am y 7:00 pm para los metales Cu y Zn, en el interior de la laguna. Se puede observar también que tanto para Cu y Zn, las concentraciones en todos los puntos de muestreo son mayores a las 7:00 am en comparación con los obtenidos a las 7:00 pm.

Las concentraciones de Cu, en la superficie (0,40 m de profundidad) son mayores que las observaciones para las muestras tomadas a 2 m de profundidad, este efecto se observa tanto en el muestreo realizado a las 7:00 am como en el de las 7:00 pm y puede atribuirse al tipo de flujo existente en la laguna facultativa, obedece al comportamiento de un flujo pistón; en el cual un 20% del volumen son espacios muertos y corto-circuitos (Aldana1 et. al., 1995). Adicionalmente influye la acción del viento la cual cambia la dirección cada 12 horas durante el día a una velocidad de 12.6 Km/h medida sobre el nivel del suelo (M.A.R.N.R, 1986).

El Zinc se remueve en un 90% de su valor inicial entrada de la laguna facultativa. En la Fig. 4 y 5 se grafican los valores de la concentración promedios a la 7:00 am y 7:00 pm para los metales Cr y Fe en el interior de la laguna. Se observa que los valores de la superficie son ligeramente superiores que a los encontrados en el fondo para las dos muestras.

Se observa además en el perfil de las 7:00 am que existe un comportamiento similar al obtenido en la distribución para el Cu y el Zn (Fig. 2 y 3) con la excepción de los puntos (3, 5 y 7), ubicados en el margen izquierdo a favor del flujo donde se aprecia que los valores para el Cr y Fe se incrementan, observándose este efecto a nivel de la superficie y el fondo. En el caso de las 7:00 pm, el perfil obtenido es de tendencia similar para todos los metales en estudio a excepción del cobre. Esta variación característica en los tres metales Cu, Zn y Fe puede atribuírsele al efecto causado por el viento, pero además contribuye el tipo de entrada y salida de alimentación del caudal existente en la laguna. (Aldana² et. al., 1995), el cual es unidimensional, superficial y resulta favorable para el descenso de los metales pesados por

gravedad a través de la longitud total de la laguna.

En la tabla 2 y 3 se indican los porcentajes de remoción para los metales pesados Cu, Zn obtenidos a la salida de la laguna A1. En las tablas 5 y 6 se indican los valores de remoción para metales de Cr y Fe. El cromo se remueve en un 73%, no siendo detectable en la mayoría de los muestreos realizados en la salida (punto 8) y en algunos puntos de la entrada (punto 1), el porcentaje de remoción, de este parámetro sólo pudo ser calculado en dos análisis. La concentración máxima permisible para riesgo es de 0,05 mg/l. El hierro se remueve en un 64% de su valor inicial, es el metal con mayor concentración presente en la laguna; sin embargo, su concentración a la salida está por debajo de los límites permisibles. La concentración máxima de hierro permitida para agua tratada y riego es de 1.0 mg/l (Gaceta Oficial República de Venezuela, 1995).

**Conclusiones**

[Evidencia] Los cuatro metales pesados (Cu, Zn, Fe, Cr) estudiados presentan un comportamiento de remoción similar, donde se pone en evidencia la influencia del tipo de mezcla, la dirección del viento, oxígeno disuelto, la temperatura y el pH, para lograr niveles de eficiencia tan altos como es el caso de la laguna facultativa A1. [Elaboración] El tipo de entrada y salida unidimensional construida en la laguna facultativa influyen en la dispersión de la concentración de los metales, [Condición] los cuales deben mantener una uniformidad en el interior de la masa de agua si se cumpliera un patrón de mezcla completa. [Elaboración] En este estudio se demuestra la influencia de un patrón de flujo pistón, el cual pudiera ser mejorado variando el número de entradas y salidas a la laguna, con la finalidad de disminuir la dispersión actualmente presente. En el sentido transversal de la laguna se genera un patrón de mezcla completa, manteniendo la oxigenación de la masa de agua durante el día influenciada por la dirección del viento. La eficiencia de remoción de los metales no está solamente condicionada por la precipitación; sino también, por la influencia de las algas y bacterias órgano-reductoras presentes durante el proceso biológico que utilizan los metales como micronutrientes y cofactores de las reacciones enzimáticas (Meiring y Ollerman, 1994).

Figura 34. Ejemplo de Artículo donde se construye una red semántica

### Análisis Semántico 1

En el texto se aprecia que los elementos semánticos que se describen en el texto. Los **ejes Temáticos** que se aprecian en el en el texto son los siguientes: Laguna Facultativa, Metales pesados, Concentración de Metales, la acción del viento, temperatura, tipo de flujo y la fotosíntesis, Lagunas Facultativas y

Laguna de Maduración, Espectrofotometría de absorción atómica y Agua residual (Figura 35).

**Ejes Semánticos:**

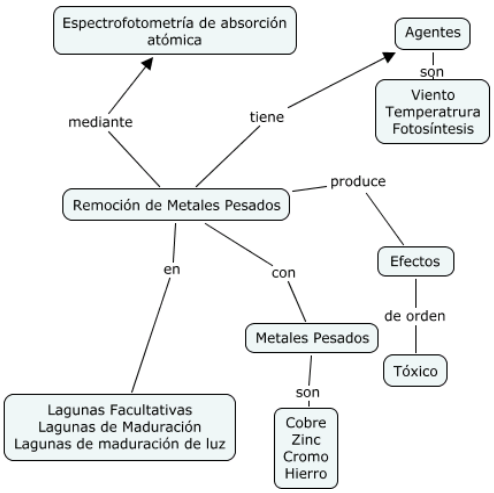



Figura 35. Ejes semánticos graficados del apartado introducción

El autor del artículo deja bien claro que su eje semántico principal es la Remoción de metales pesados, para ello describe los agentes que influyen en el proceso, declara que serán los metales pesados sobre los que trabajará, las técnicas que empleará para la remoción de estos metales es la Espectrofotometría de absorción atómica y el contexto de aplicación del experimento son las Lagunas de Maduración y las Lagunas Facultativas, por último declara los efectos de este proceso contaminante (efectos toxicológicos). Estos elementos son llevados a la ontología y permiten organizar la estructura con las dimensiones de los ejes semánticos. Los términos que se extraen con la herramienta Word Smith Toolls son llevados a la base de datos terminológicos para su tratamiento lexicográfico con vistas a ser posteriormente usados en la ontología, además las claves de análisis facilitan organizar los segmentos de términos bajo las cláusulas siguientes: Procesos, Aplicación, Efectos y Técnicas de Medición.

El estudio de los ejes semánticos arrojó lo siguiente (Ver tabla 15). Esto evidencia que los ejes semánticos que más se describen en el estudio son: Procesos, Agentes, formas de medición, aplicaciones de algunas actividades,

los efectos de algún fenómeno sobre otro, las técnicas de representación, etc. En el estudio también se localizaron otros ejes de menos nivel como instituciones públicas (Anexo 6).

| Variable | Frecuencia |
|---|---|
| Procesos | 80 % |
| Agentes | 96 % |
| Medición | 50 % |
| Aplicación | 70 % |
| Efectos | 98 % |
| Técnicas | 93 % |
| Representación | 99 % |

Tabla 15 Ejes Semánticos del dominio Ingeniería de Puertos y Costas.

### 5.4.2.- Estilo y Retórica

En el apartado el estilo y la retórica del discurso científico se manifiesta de forma variada. El autor es capaz de declarar los derroteros de la investigación a través de elementos que sugieren pormenorización, entre ellos se puede observar temperatura, flujo, métodos de investigación como la fotosíntesis. El autor declara sustantivos combinados con adjetivos relativos al dominio de la Química. También expone el autor en su discurso retórico elementos característicos de la actividad científica, datos factuales como profundidad, así como la manifestación de los métodos para realizar la investigación. Con estos elementos el autor hace creíble su discurso y sigue la misma retórica asociada al medio de las ciencias, explicitando cómo, dónde, cuándo y qué efectos se obtuvieron de la investigación.

### 5.4.3.-Extracción de Términos (Microestructura)

En esta fase se recogen los términos para formar una lista plana, donde ningún término tiene más valor semántico que otro. Seguidamente se elaboró una lista con los conceptos más importantes, amparados en los criterios de los especialistas y de la herramienta de análisis léxico Wordsmith Tools. Finalmente se construyó una lista de frecuencias que permitió inferir el conocimiento especializado en el domino de la Ingeniería de Puertos y Costas. La identificación de las palabras clave facilita el modelado conceptual de la ontología que articulará todo el conocimiento en la aplicación. Los lemas que

representa a los vocablos más frecuentes admiten la identificación de las diversas categorías conceptuales sobre las que se basa la definición de los términos del texto. La figura (figura 36) muestra el resultado del análisis del corpus en español realizado por Wordsmith Tools a partir de los documentos extraídos tras la primera fase del proyecto (Ver figura 36).

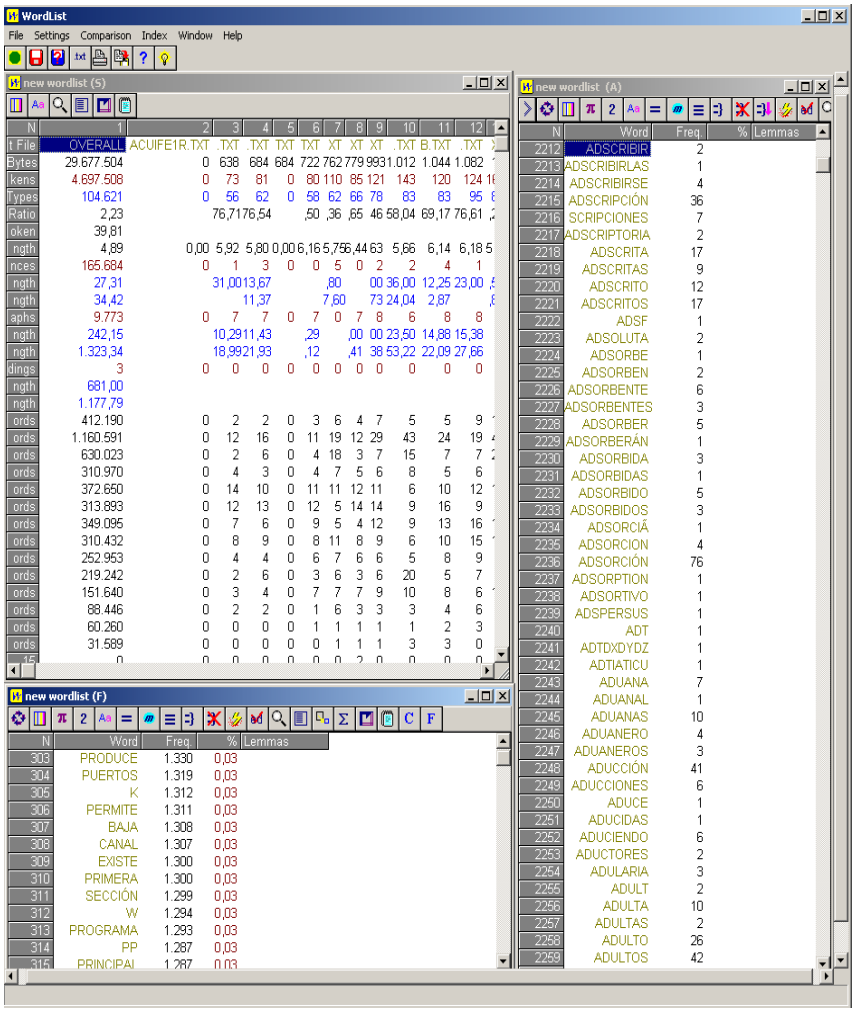



Figura. 36 Resultado mostrado por Wordsmith Tools
del análisis de los ficheros de texto (Senso, 2007).

Para lograr que cada término esté dentro de su uso correcto, se usaron concordancias, es decir, se apeló a la exposición de las ocurrencias de determinado vocablo en su contexto lingüístico, de tal forma que la palabra en cuestión aparece en el centro de una línea y a ambos lados se muestran las palabras con las que suele aparecer en los textos. Como apunta Senso (2007)

se trata de una técnica muy similar a la empleada en los índices KWIC (Keyword In Context) pero con una finalidad diferente, ya que aquí se emplea menos para recuperar y más para contextualizar.

### 5.4.3.- Macroestructura

El análisis macroestructural nos aporta específicamente los elementos globales que caracterizan a la estructura general del texto (Anexo 5). Los resultados del estudio macro estructural se declaran en las siguientes tablas (Ver Tabla 16).

| Macroestructura | % |
|---|---|
| Introducción | 100 % |
| Metodología | 96 % |
| Objetivos | 67 % |
| Resultados | 86 % |
| Conclusiones | 100 % |

Tabla 16: Elementos Macro-estructrurales del Corpus

Los resultados del análisis de la macroestructura global del texto, declaran que los textos tienen en un 100 % el apartado introducción, en un 96 % metodología y en un 100 % conclusiones. Los apartados de más baja aparición en los textos son objetivos y conclusiones, esto se obtiene debido a la heterogeneidad del corpus de estudio.

Las oraciones de Síntesis se declaran de la siguiente forma: La mayoría de los textos analizados poseen oraciones de síntesis al final de cada párrafo, el 25 % de los textos posee las oraciones de síntesis en la parte intermedia del texto. Solo el 20 % posee estas oraciones aparece en la posición inicial y el 5 % de los textos posee las oraciones de cierre conclusivo al final del párrafo Conclusiones (Ver tabla 16).

| Oraciones de Síntesis | Posición final | Posición intermedia | Posición inicial | Cierre conclusivo |
|---|---|---|---|---|
| % | 60 % | 25 % | 20 % | 5 % |

Tabla 16: Elementos Macro-estructrurales del Corpus

Si se analiza el nivel de síntesis pueden suprimirse oraciones que poseen bajo nivel semántico y bajo nivel de generalización. Según la tabla existen oraciones

de bajo nivel semántico en un 25 % de los textos y de bajo nivel de generalización en un 18 % de los documentos, esto dice mucho de la baja calidad de redacción de los documentos (Tabla 17).

| Oraciones de Síntesis | Bajo Nivel Semántico | Bajo Nivel de Generalización |
|---|---|---|
| % | 25 % | 18 % |

Tabla 17: Estructura de bajo nivel semántico en oraciones de síntesis

### *5.5.- Análisis del corpus*

El modelo de resumen automático que proponemos se desarrolla a partir del análisis previo de las estructuras lingüísticas de 50 artículos de revistas de Ingeniería de Puertos y Costas y de sus correspondientes 50 resúmenes, los cuales conforman nuestro corpus de referencia.

De los 50 textos que se han seleccionado para la investigación se analiza la estructura textual, las unidades léxicas representativas, la estructura discursiva y la sintáctico-comunicativa de los textos. Teniendo en cuenta el análisis de discurso y el estudio de los elementos estilísticos, retóricos y lingüísticos de los corpus se proponen reglas de carácter cohesivo, estructural y pragmático que facilitan la construcción de regularidades o lineamientos para implementar la herramienta de extracción de corpus.

#### 5.5.1-Tipología de reglas para el desarrollo del resumen

Las reglas que han de emplearse en el desarrollo del resumen, son de diversa naturaleza y tratan de adecuar los resultados teóricos y los análisis textuales al desarrollo de la aplicación, las mismas han sido elaboradas a partir del estudio del corpus y constatadas en los resúmenes desarrollados por expertos (Anexo 12):

1. Identificando unidades léxicas que indican la supresión de segmentos textuales del texto original, en su totalidad, o secciones oracionales.
2. A partir de lo expresado en el punto uno, se proponen reglas que eliminan oraciones, es decir, se proponen patrones para suprimir determinadas oraciones o segmentos textuales, determinado de esta

forma posiciones sintáctico-comunicativas y discursivas.

3. Teniendo en cuenta los enunciados textuales, es posible generar lineamientos que regulen o aumenten la ponderación de las oraciones en el texto original.
4. Mediante la observación de los procesos de desambiguación y corrección de textos es posible construir un mecanismo de desambiguación que facilite que la base de datos (ontología) pueda generar acciones para mejorar la cohesión léxica en cada documento.

### 5.5.2- Análisis de la estructura textual del artículo científico en el dominio Ingeniería de Puertos y Costas

En este apartado exponemos el análisis textual que se realiza sobre los textos del subcorpus de referencia. Si bien en los apartados anteriores fuimos desde la parte exterior del corpus, este análisis es hacia adentro, buscando lo que no se declara, pero connota semántica y sintácticamente en los resultados. En primer lugar, se lleva a cabo un análisis de la estructura textual de los artículos y de los resúmenes, a partir de su macroestructura, con el objeto de demostrar la coincidencia retórica discursiva I (O) MRC. El otro análisis que se realiza en este segmento de la investigación consiste en un análisis textual (retórico) de cada apartado de cada artículo de la muestra, para determinar, la existencia de bloques oracionales con determinada carga semántica, cuya relevancia pueda ser imprescindible para el resumen futuro.

#### 5.5.2.1- Análisis de los subtítulos del artículo científico en Ingeniería de Puertos y Costas: estructura I(O) MRC

El análisis de los subtítulos de los artículos especializados en Ingeniería de Puertos y Costas tiene la finalidad de constar la existencia de una estructura I(O) MRC (Introducción, Metodología, Resultados y Conclusiones). Las variantes de los subtítulos de los textos seleccionados pueden verse reflejados en la siguiente tabla (Tabla 18).

| Apartado | Título | Frecuencia T. | Frecuencia R. |
|---|---|---|---|
| 1 | Fundamento | 10 | 34 |
| | Objetivos | 24 | 26 |
| 2 | Metodología | 26 | 26 |
| | Métodos | 11 | 8 |
| | Material y Métodos | 10 | 12 |
| | Población y Muestra | 3 | 3 |
| 3 | Resultados | 50 | 50 |
| | Resultados -Discusión | | |
| 4 | Conclusiones | 44 | 44 |
| | Conclusiones y Discusión | 6 | 6 |

Tabla 18: Variación de la estructura de las Secciones de los Textos del Dominio

En los artículos se observa una línea homogénea en el tratamiento del título. En todos los apartados se aprecia la presencia de un título sin variación en los parágrafos, tanto en el apartado 3 (Resultados), como en el 4 (Conclusión). En cambio, el título del apartado 2 (Metodología), es el que más variedad de presentación presenta, lo mismo puede verse Metodología, Material y Métodos, ,Población y Muestra, etc., siendo las ocurrencias que presentan más variedad en sus formas, aunque la representación más evidente de esta es Metodología con 26 ocurrencias.

Los apartados 3 y 4 no evidencian variaciones significativas. Lo que diferencia estos apartados de los resúmenes de los autores son los subtítulos de los artículos y los subtítulos de los resúmenes , pues en estos se ofrecen varias denominaciones, siendo fundamento con 44 ocurrencias, Fundamentos, Objetivo y Objetivos, las formas de subtítulos más utilizadas para denotar el apartado 2 en los resúmenes de los artículos. Sin embargo, aunque existe gran variedad de denominaciones en esta estructura, ha sido Metodología (con 52 ocurrencias) la forma que más se usa para este segmento de artículo científico.

Mediante este análisis constatamos que los autores de artículos especializados

en Ingeniería de Puertos y Costas siguen una retórica en la composición y construcción de los textos, que tiene sus diferencias en el tratamiento del resumen.

#### 5.5.2.3.- Análisis de los apartados del artículo en Ingeniería de Puertos y Costas

El análisis de los apartados del artículo en Ingeniería de Puertos y Costas se sustenta en los criterios desarrollados en Edmunson (Edmundson, 1969) y D´Cunha (D'Cunha, 2006), donde se apela a la posición de determinados vocablos en las oraciones como indicadores de la relevancia de los segmentos de texto, lo que evidencia la presencia de cargas semánticas en aquellas oraciones del texto donde se repiten diversos términos. Los procederes en que se apoya nuestro análisis se centran en los siguientes criterios:

1 Es evidente la existencia de segmentos textuales donde la información posee mayor nivel semiótico.

2 Existen zonas donde existen oraciones cuyo nivel de relevancia temática e informativa supera otras.

3 La unidad semántica textual Introducción describe el Objetivo del artículo y su nivel semántico esencial está declarados en las tres primeras oraciones del párrafo inicial y en las 4 últimas oraciones del párrafo final.

4 El apartado Metodología declara sus valores semánticos en las primeras tres oraciones y en las 4 últimas.

5 En el parágrafo Resultados los tres última oraciones son los de mayor carga semántica

6 En la sección Conclusiones se da un alto nivel de carga semántica y relevancia informativa, sus primeras 5 oraciones y las 3 finales poseen toda la información necesaria del apartado.

Observamos que en otras posiciones de los artículos también hay contenidos destacados, pero que entre las oraciones que se encuentran en las posiciones mencionadas, siempre hay alguna que contiene información relevante (Anexo 6).

De este análisis se evidencia que existen determinadas oraciones cuya carga semántica está ubicada en determinados puntos en el párrafo, lo que demuestra la existencia de posibles sentencias candidatas para el resumen de acuerdo a la relevancia semántico-estructural a nivel oracional (ver Tabla 17).

| Elemento estructural | Posición de la fuerza semántica |
|---|---|
| Introducción | 3 primeras oraciones y 4 últimas |
| Metodología | 3 primeras y 3 últimas oraciones |
| Resultados | 3 últimas oraciones |
| Conclusiones | 5 primeras o 3 últimas oraciones |

Tabla 19. Elementos estructurales de mayor fuerza semántica

Es evidente que estas ocurrencias no son una receta, por tanto su frecuencia de aparición en los artículos no es definitiva. Los trabajos de Edmunson (1969) y Hovy y Lin (1997, 1999) y los postulados de (Arco, 2008) prueban que los procesos con texto se desarrollan en ocurrencias observadas en más de 5 ocasiones en determinados textos.

Es importante reconocer que los procesos de ocurrencias de eventos que se describen aquí, son elementos que eran básicamente desconocidos por el autor, de lo contrario se habrían identificado con claridad estos dispositivos textuales para facilitar el resumen, al menos por extracción de oraciones.

Otro aspecto importante dentro de este estudio es verificar la relevancia de determinados segmentos textuales dentro del texto. Para realizar este análisis se manutuvo la misma muestra de 50 artículos (Ver tabla 20) y se siguió el proceder que sigue a continuación:

1 Seleccionar los 50 artículos.

2 Extraer las oraciones situadas en los apartados de los artículos que se van a analizar y el número de palabras que los contienen.

3 Confrontar las oraciones con las del resumen del autor, comprobando cuales expresan los mismos contenidos y cuáles no, y en qué apartados están.

4 Declarar la relevancia de las oraciones encontradas en las zonas

indicadas en el artículo.

| No.Art. | Título | No. de palabras | O/OB |
|---|---|---|---|
| 1 | Modelado en clima árido | 12345 | 0 |
| 2 | La propiedad de aguas perennes en el sureste ibérico | 1345 | 0 |
| 3 | El Agua | 2134 | 0 |
| 4 | Laguna Mata Redonda | 1567 | 0 |
| 5 | Contaminación Provocada por los Sedimentos | 1340 | 0 |
| 6 | Meteorización | 2167 | 0 |
| 7 | Hidráulica | 3456 | 0 |
| 8 | Abastecimiento de agua | 1789 | 0 |
| 9 | Acequias y aljibes | 1098 | 0 |
| 10 | Uso del agua | 1870 | 0 |
| 11 | Características del agua | 1319 | 0 |
| 12 | Acuíferos semiconfinados | 2368 | 0 |
| 13 | La instalación piloto de recarga artificial de “los sotillos” | 4568 | 0 |
| 14 | Modelos geológicos en los acuíferos kársticos del norte de la provincia de Málaga: implicaciones hidrogeológicas | 2987 | 0 |
| 15 | Primera aproximación mediante modelización al análisis de la influencia del embalse de rules en el régimen hidrológico del acuífero de Motril-salobreña (Granada) | 1345 | 0 |
| | | 2789 | 0 |
| | | | 0 |
| 16 | El conocimiento de los acuíferos del campo de dalias y su implicación en la gestión sostenible integral de los mismos | 2134 | 0 |
| 17 | Simulación de alternativas de aprovechamiento hídrico en el | 5689 | 0 |

| | | | |
|---|---|---|---|
| | acuífero de Guadix-marquesado tras el cierre de la mina de Alquife | | |
| 18 | Calidad agua | 3467 | 0 |
| 19 | Calidad de los materiales | 1200 | 0 |
| 20 | Conservación del agua | 1234 | 0 |
| 21 | Micromedición | 5678 | 0 |
| 22 | Hidráulica Sistemas de conducción | 1560 | 0 |
| 23 | Recurso agua | 1876 | 0 |
| 24 | Historia del uso del agua | 1981 | 0 |
| 25 | Aplicaciones de la Simulación Hidrológica en Zonas Áridas | 1976 | 0 |
| 26 | Principales Contaminantes | | 0 |
| 27 | Las cuencas sedimentarias | | 0 |
| 28 | Producción de Alimentos e Impacto Ambiental | 1845 | 0 |
| 29 | Modos de desplazamiento por elementos | 1930 | 0 |
| 30 | Agua: estado natural, humedad relativa, capilaridad, hidrometeoro, ciclo biogeoquímico | 1835 | 0 |
| 31 | Agua: recurso natural. Acuicultura, Piscicultura | 1564 | 0 |
| 32 | Gestión del agua | 1783 | 0 |
| 33 | Ventajas y limitaciones del uso del mercado en la asignación de los Recursos Hídricos | 1873 | 0 |
| 34 | Nuevas Herramientas Tecnológicas como Apoyo a los Sistemas de Gestión. | 400 | 0 |
| 35 | Instrumentos económicos para la gestión del agua en América Latina y el Caribe: el caso del mercado del agua en Chile | 519 | 0 |
| 36 | Predicción de la Erosión de Suelos | 1615 | 0 |
| 37 | Hidrología y Geodinámica de la Cuenca Amazónica | 1256 | 0 |
| 38 | Depuración y reutilización de aguas residuales | 2789 | 0 |
| 39 | El agua subterránea: calidad y | 3531 | 0 |

| | | | |
|---|---|---|---|
| | contaminación | | |
| 40 | Arenas residuales | 3456 | 0 |
| 41 | Análisis de precipitaciones anuales | 2345 | 0 |
| 42 | Formas geomorfológicas erosivas, caudales sólidos y líquidos | 4676 | 0 |
| 43 | Erosión y transporte de sedimentos | 2134 | 0 |
| 44 | Vertederos de excedencia | 3456 | 0 |
| 45 | Inestabilidad geomorfológica | 1235 | 0 |
| 46 | Intensidad del fenómeno el Niño | 1567 | 0 |
| 47 | Límites entre mar y continente | 1872 | 0’ |
| 48 | Formación de las precipitaciones | 2478 | 0 |
| 49 | Estructura vertical de la atmósfera | 1732 | 0<br>0 |
| 50 | Procesos elementales de erosión: la meteorización | 1567 | 0 |

Tabla 20: Cantidad de Caracteres existentes en los artículos

En la tabla se muestra la cantidad de términos en determinados segmentos oracionales en los artículos del corpus de referencia, donde pueden verse los de mayor relevancia en el artículo.

Consideramos que esta investigación textual nos está completa si no se declara la posición que presentan determinados segmentos de texto en el resumen, para ello se ha decidido utilizar 10 artículos de la muestra de corpus debidamente desarrollada en la investigación.

| Ref. | Oración | Incluido | No |
|---|---|---|---|
| 73ap.1 | El uso eficiente del agua y el reúso de grandes volúmenes de agua residual<br><br>municipal tratada proveen el abastecimiento de agua para un amplio rango de<br><br>propósitos municipales, industriales, agrícolas y recreativos. | | X |
| 74ap.1 | A las estructuras que se construyen para promover la infiltración de agua | Recientemente en Japón el sistema de infiltración de agua | |

| | | | |
|---|---|---|---|
| | pluvial por medios artificiales se les da el término de instalaciones de infiltración. | pluvial se está convirtiendo gradualmente en una parte integral de las medidas para la preservación del ciclo hidrológico en áreas urbanizadas. | |
| 75ap2 | La configuración topográfica de la Cuenca unida a un régimen de lluvias temporal<br><br>e irregular hacen poco viable, técnica y económicamente, un incremento de la<br><br>regulación por los procedimientos tradicionales (presa y embalse). | La Cuenca del Guadalquivir, con una extensión del orden de 60.000 Km2, ocupa el cuarto lugar entre los grandes ríos españoles. Su característica fundamental es la extremada irregularidad de su régimen, que hace que sus caudales, en régimen natural, puedan oscilar, a lo largo del año, en la proporción de 1 a 1.000 y sus recursos brutos, entre el año más seco y el más lluvioso, en la proporción de 1 a 5, siendo los correspondientes al año medio, del orden de 8.900 Hm3. | |
| 76ap2 | Los ingenios azucareros consumen cantidades considerables de agua durante el<br><br>procesamiento de la caña de azúcar. | | X |
| 77ap2 | Los objetivos que cumplen los estudios hidrológicos en proyectos que utilizan los recursos hidráulicos de una cuenca son los mismos en un desarrollo pequeño que en uno grande. | Los principales objetivos de los estudios hidrológicos. | |
| 50ap2 | Los experimentos consisten en el vertido de trazadores lagrangianos y la filmación de su comportamiento. | Los experimentos desarrollados para la realización de la investigación consisten en el vertido de trazadores lagranjianos. | |
| 60ap2 | La metodología y los experimentos (que incluyen casos de oleaje regular, oleaje irregular y reproducción de oleajes registrados en prototipo) han | La concepción metodológica de este proyecto fue realizada mediante la detección de casos de oleaje regular. | |

| | | | |
|---|---|---|---|
| | sido los mismos para ambos laboratorios. Sin embargo, los resultados no son tan semejantes como hubiera sido de esperar. | | |
| 63 ap3 | La variación de la cota relativa de la draga respecto al tiempo (draga llena y draga vacía), con el registro del momento de máximo llenado (relacionado con la capacidad de transporte o con el volumen dragado) y del momento de vaciado (relacionado con el punto de vertido). | Como resultado se obtiene que la variación relativa de la cota de acuerdo a condiciones determinadas. | |
| 64 ap4 | En este trabajo se presentará los últimos avances realizados en el estudio del proceso de rotura mediante la utilización de modelos que resuelven las ecuaciones generales en el dominio del tiempo. | Se presentan los últimos avances obtenidos en el estudio de las roturas mediante el empleo de ecuaciones generales como modelos matemáticos. | |
| 48 ap3 | Este trabajo demuestra la importancia de las resonancias debidas a ondas de bordes. Los resultados preliminares pueden encontrarse en Ciriano et al, (2000). | Se derribe el valor de las resonancias originadas mediante ondas de bordes. | |

Tabla 21. Fragmentos del texto usados en el resumen.

En la tabla anterior puede apreciarse la aparición de oraciones relevantes del texto en determinadas secciones del texto, esto evidencia la existencia de determinadas oraciones de relevancia para el resumen.

A continuación se muestran algunos elementos que demuestran la relevancia posicional de algunas oraciones en los 4 apartados del texto (IOMRC) (Tabla 22).

| Texto | Aptado 1 | Aptado.2 | Aptado.3 | Aptado.4 | Total |
|---|---|---|---|---|---|
| 73ap1 | 1.or | 1.or. | 2.or. | 3.or. | 7.or. |
| 74ap1 | 2.or | 1.or. | 2.or. | 4.or. | 9.or. |
| 75ap2 | 3.or | 2.or. | 2.or. | 4.or. | 10.or. |
| 76ap2 | 1.or. | 2.or. | 1.or. | 1.or. | 6.or. |
| 77ap2 | 1.or. | 2.or. | 1.or. | 1.or. | 5.or. |
| 50ap2 | 1.or. | 2.or. | 3.or. | 1.or. | 7.or. |
| 60ap2 | 1.or. | 1.or. | 3.or. | 2.or. | 7.or. |
| 63ap3 | 1.or. | 1.or. | 3.or. | 2.or. | 7.or. |
| 64ap4 | 1.or. | 1.or. | 3.or. | 4.or. | 9.or. |
| 48ap3 | 2.or. | 1.or. | 3.or. | 4.or. | 10.or. |
| **Total** | 14 | 14 | 23 | 26 | 77 |

Tabla: 22. Cantidad de Oraciones en cada Apartado del texto

Los resultados de la coincidencia de oraciones en los apartados manifiestan que:

1. Son 14 las oraciones que aparecen entre las tres finales del segmento Introducción que son declaradas en el resumen de los atores siendo el 19 % de los contenidos que se declaran por los autores en sus sumarios.
2. 14 oraciones están situadas entre las dos primeras en el parágrafo Metodología, con un 19 % de la muestra.
3. 29 % de las oraciones del apartado Resultado aparecen utilizadas por los autores en sus resúmenes.
4. De las 10 oraciones declaradas entre las primeras y terceras del apartado Conclusiones, lo que representa 38 % para este apartado, una carga importante de información para el resumen.

Es evidente el nivel de incidencia semántica y estructural de las oraciones en los apartados esenciales del estudio, pues 61 de las 77 oraciones (79%) aparecen situadas en los acápites Introducción (Objetivos), Metodología, Resultados y Conclusiones.

A continuación en la figura se ejemplifica el análisis de un texto del corpus de

artículos. El título del artículo aparece en mayúscula y seguidamente se muestra el número de oraciones de su resumen.

En el ejemplo también aparece la cantidad de oraciones que no poseen una referencia directa en el artículo base, a través de sentencias que poseen los verbos declarados y la cifra de oraciones que están referenciadas por unidades oracionales, en las que existen algunos de los verbos anteriormente referidos.

La estructura retórica del resumen aparece en negrita con sus oraciones declaradas en letra Verdana 10. Según D´Cunha (2006) en estas oraciones se analiza la presencia de oraciones similares a las del artículo (subrayadas) de haberla(s), las oraciones del artículo (en negrita) que contengan alguno de los verbos de la lista (subrayados) pero que además aporten contenidos incluidos en las oraciones del resumen del autor".

El autor ha tenido en cuenta la posible existencia de un número considerable de oraciones del artículo que hagan referencia a una o varios grupos oracionales en el resumen, fenómeno de composición que se da por las siguientes causas:

1 Repetición de ideas en el texto mediante dos sentencias u oraciones.

2 Grupos oracionales excesivamente largos que generalmente se seccionan y que aporta siempre una misma información (separación incorrecta de los grupos de oraciones).

3 La presencia de oraciones con varias unidades léxicas o léxico-semánticas.

El resumen que se ha analizado posee 11 oraciones, 8 de ellas aparecen reflejadas en el artículo que se analiza, además, existen verbos de la lista que están contenidos en disímiles unidades de sentido (oraciones) del artículo contentivo de alguno de los verbos declarados en la lista.

Se ha declarado en las oraciones que están subrayadas los verbos que aparecen en la lista. Es evidente que existen tres oraciones que no poseen referencia mediante verbos de la lista. En esta figura aparece un listado de verbos (con su frecuencia de aparición en el texto) declarados en las unidades

oracionales del texto científico que están en consonancia semántica con las oraciones utilizadas por los autores para su resumen (Ver Tabla 23).

| Texto Tipo |
| --- |
| Título: RESONANCIA EN UN PUERTO FRENTE A ONDAS DE BORDE.<br><br>11 oraciones en el resumen, 3 no referenciadas en el artículo científico con estos verbos y 8 referenciadas mediante estos verbos que se detallan a continuación:<br><br>atrapar 1<br><br>tener 1<br><br>generar 1<br><br>propagar 2<br><br>originar 1<br><br>estudiar 1<br><br>contribuir 1<br><br>interrumpir 1<br><br>realizar 2<br><br>transmitir 2<br><br>enviar 1<br><br>reflejar 2<br><br>delimitar 1<br><br>**Introducción**<br><br>Las ondas de borde son ondas gravitatorias ***atrapadas*** en la costa debido a la refracción topográfica (Ursell, 1952). Algunos puertos españoles, especialmente en la costa Mediterránea, ***tienen*** su bocana orientada hacia largas playas donde, fácilmente, las ondas de borde se pueden ***generar*** y ***propagar***. Al incidir estas ondas en el puerto se pueden ***originar*** efectos |

resonantes. Una de las finalidades de este trabajo es ***estudiar*** dichos efectos.

Por otro lado, la mayoría de los estudios sobre la propagación de ondas de borde se ***realizan*** sobre una topografía uniforme en una playa rectilínea. Este trabajo pretende ***contribuir*** a la investigación del efecto de una ***interrupción*** topográfica sobre la propagación de estas ondas. Un problema similar ha sido investigado por Santos y Peregrine, 1998.

**Metodología**

Se ***considera*** una geometría idealizada como la de la figura 1. Se trata de un puerto rectangular conectado a una playa. Se ***envía*** una onda de borde monocromática hacia la bocana del puerto. Debido a la variación topográfica, parte de la energía enviada se ***transmite*** y parte se ***refleja***. La energía ***reflejada*** se ***propaga*** en forma de ondas de borde de diferentes modos y otros tipos de ondas, algunas de ellas no quedan atrapadas. Con la finalidad de ***evitar*** las dificultades originadas por estas ondas se ***delimita*** el dominio mediante un contorno artificial (pared vertical) paralelo a la costa.

A partir de las ecuaciones de conservación de masa y momento bajo la aproximación para aguas someras se ***realiza*** un estudio parcialmente analítico y parcialmente numérico del sistema

Tabla.23.Ejemplo de verbos utilizados en un texto

Después de terminar el análisis de los 10 artículos con sus respectivos resúmenes, se contabilizan los resultados. Hay 201 oraciones en todos los resúmenes, de las cuales 13 son oraciones usadas siempre en las 51 oraciones de los resúmenes, 162 de ellas (el 80 %) aparece con nexos referenciales ya sea mediante paráfrasis o por repetición textual en oraciones que poseen verbos de la lista, solo un 20 % no aparecen referenciadas en oraciones que contengan verbos listados.

Los verbos de mayor frecuencia se detallan en las tablas (Ver tabla 24 y 25) que se describen a continuación, en las mismas pueden verse que las formas verbales más utilizadas son: analizar, determinar, evitar. Los verbos de menor frecuencia de aparición son atrapar, y propagar, además se declaran las frecuencias de los verbos en determinados textos, siendo el texto dos el de más riqueza verbal en la muestra analizada. Esta situación permite conocer la coincidencia entre los verbos de la lista y su número correspondiente de

oraciones en consonancia con los contenidos que aparecen en las oraciones.

| Textos | T.Resúmenes | No de Oraciones1 | No de Oraciones 2 |
|---|---|---|---|
| Texto 1 | 18 | 18 | 8 |
| Texto 2 | 20 | 17 | 6 |
| Texto 3 | 17 | 15 | 8 |
| Texto 4 | 20 | 19 | 6 |
| Texto 5 | 19 | 16 | 7 |
| Texto 6 | 15 | 14 | 8 |
| Texto 7 | 16 | 12 | 6 |
| Texto 8 | 20 | 18 | 7 |
| Texto 9 | 19 | 19 | 7 |
| Texto 10 | 17 | 16 | 8 |
| Total | 201 | 162 | 69 |

Tabla.24. Frecuencia de verbos en determinadas oraciones del texto y en el resumen

| Verbos de la lista | F. | T.1 | T.2 | T.3 | T.4 | T.5 | T.6 | T.7 | T.8 | T.9 | T.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| analizar | 10 | | 1 | 3 | 5 | | 1 | 1 | | | |
| atrapar | 2 | | 1 | | 1 | | | | | | |
| tener | 1 | | | 1 | | | | | | | |
| generar | 8 | | 4 | | 1 | | 1 | | 2 | | |
| propagar | 2 | | | | | | | | | 2 | |
| originar | 3 | | | | | 2 | | | | | |
| estudiar | 9 | 2 | | | 2 | | | 3 | 2 | | |
| contribuir | 5 | | 3 | | | 1 | | | | | 2 |
| interrumpir | 7 | | 1 | 2 | | | 2 | | | | 1 |
| realizar | 9 | | 1 | 1 | 1 | 2 | | 1 | 1 | 2 | |
| transmitir | 6 | 2 | | | 2 | | 1 | | | | |
| evitar | 8 | | 1 | 1 | 2 | 3 | | | | 1 | |
| reflejar | 7 | | 1 | 1 | 2 | | 1 | | 2 | | |
| delimitar | 7 | 2 | 1 | | | | | | | | 4 |
| conocer | 9 | | | | | 3 | 3 | 1 | 2 | 1 | |
| evaluar | 6 | | | 2 | 1 | 2 | 1 | | | | |
| determinar | 8 | 2 | 1 | 2 | | | 1 | 1 | 1 | | |
| incluir | 3 | | | | 2 | 1 | | | | | |

Tabla.25. Repetición de los verbos en determinados textos

## *5.6.- Análisis de las unidades léxicas representativas del artículo científico en la ingeniería de Puertos y Costas*

### *5.6.1.-Desarrollo de reglas basadas en unidades léxicas*

Este análisis tiene como objeto el reconocimiento de las oraciones de los artículos para detectar aquellas unidades léxico-semánticas, que son susceptibles de incluir o eliminar a la hora de redactar sus resúmenes. A partir de las observancias halladas en los textos estudiados se formulan reglas acorde a unidades léxicas (indicadoras de información relevante y unidades léxicas indicadoras de información irrelevante) (Anexo 4).

De acuerdo a la metodología desarrollada por D´Cunha (2006) se desarrolla un análisis de los textos seleccionados para la investigación. En este análisis se intenta detectar las unidades léxicas que subyacen en los textos especializados en Ingeniería de Puertos y Costas, pues los mismos son indicios de la relevancia de determinada oración o sentencia en el momento de ser seleccionadas para el resumen. A la vez, se localizan aquellas unidades cuyo valor permita eliminar las oraciones que poseen bajo nivel semántico de cara el resumen automático.

Al igual que en los trabajos de D´Cunha, (2006), Álvarez de Mon (Alvarez de Mon, 1999) Ono (Ono et al., 1994) se intenta determinar la relevancia funcional, verbal, estructural y gráfica de de las diversa unidades léxicas. Los tipos de unidades léxicas que se identifican son:

1. Unidades léxicas nominales
2. Unidades léxicas verbales.
3. Unidades léxicas del título principal

#### 5.6.1.- Unidades léxicas que indican relevancia

Para este análisis se han detectado una cantidad de unidades nominales que representan a los 50 artículos de Ingeniería de Puertos y Costas. Este análisis textual facilita la detección de diversas unidades léxicas nominales que podrían ser facilitadoras del resumen (Anexo 4 y 17). Las unidades seleccionadas son: realiza, presenta, se ha considerado, cabe destacar, descrita, fin, estudia,

,resultados, comparación, resultado, validación, construido, analizar objetivo, objeto, propósito, método, intención, conclusión, resultado (Ver Tabla 26).

| Unidades Léxicas Nominales |
|---|
| realiza |
| presenta |
| cabe destacar |
| descrita |
| fin |
| estudia |
| resultados |
| comparación |
| validación |
| objetivo |
| objeto |
| propósito |
| método |
| intención |
| conclusión |
| resultado |

Tabla 26. Selección de verbos que indican relevancia

Estas unidades han sido detectadas mediante WorldList en un corpus de Referencia de 50 artículos en castellano. Se reconoce que hubiese sido más fácil el uso de un software de mayor alcance para este análisis, pero los datos que se logran con Worldlist son válidos. Los resultados de las listas arrojadas por el Worldlist, fueron explorados visualmente por 2 expertos con competencias gramaticales, los cuales localizaron:

1. Nombres comunes relativos a fórmulas, compuestos químicos y topónimos (esto permite controlar la sinonimia y la homonimia).
2. Unidades lexicales de acuerdo a los lemas principales para localizar vocablos en singular y plural.
3. Contextos oracionales y de párrafos (selección de frases y combinaciones (permite valorar la relevancia de oracional y de párrafo).

4. Sustantivos, adjetivos, y sus construcciones con el objeto de localizar su presencia en el texto.

En la figura (figura 36) se muestra un listado de vocablos seleccionados a partir de World Smith Toolls donde los investigadores seleccionan topónimos, nombres comunes, fórmulas y compuestos químicos.

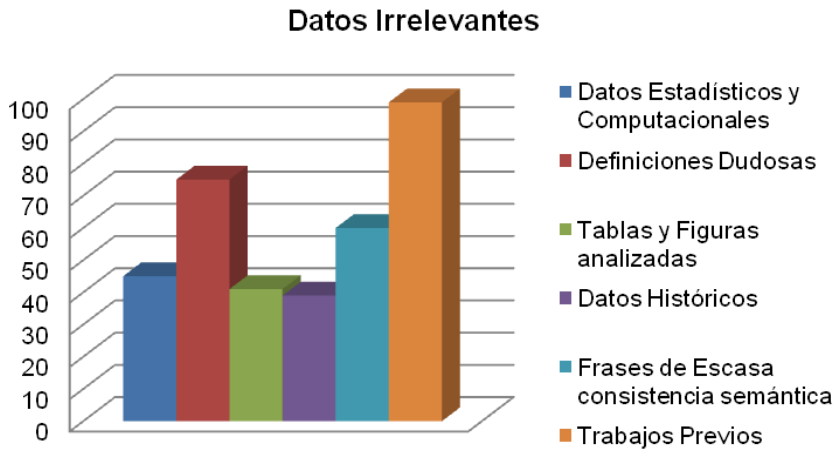



Figura 36. Cantidad de elementos que indican irrelevancia

En otros estudios como los de D´Cunha (2006) este proceso es innecesario, pues como se basa en el estudio de los resúmenes de los autores, no tiene en cuenta que en el resumen informativo se construyen oraciones de síntesis que describen específicamente resultados, objetivos, metodología y conclusiones.

Otro segmento del estudio estuvo dirigido a la detección textual de unidades léxicas de naturaleza verbal, elementos que declaran la relevancia de diversas oraciones en el texto científico. En la tabla) (ver tabla 27) se muestra un listado ordenado de los verbos incluidos en los textos especializados que componen nuestro subcorpus. Al igual que en los estudios de D´Cunha (2006), de este análisis, se exoneran aquellos verbos que son auxiliares o poco específicos como: ser, estar, haber, hacer, tener, poder y dar. Todas las formas restantes formas verbales son aceptadas, sean conjugadas o no, exceptuando el participio (por ser ambiguo muchas veces con ciertos adjetivos).

| Verbos | Frecuencia |
|---|---|
| analizar | 10 |
| atrapar | 2 |
| tener | 5 |
| generar | 8 |
| propagar | 2 |
| originar | 4 |
| estudiar | 9 |
| contribuir | 5 |
| interrumpir | 7 |
| realizar | 9 |
| transmitir | 6 |
| evitar | 8 |
| reflejar | 7 |
| | |
| delimitar | 7 |
| conocer | 9 |
| evaluar | 6 |
| determinar | 8 |
| incluir | 4 |

Tabla 27.Frecuencia de aparición de los verbos en el corpus

Del listado generado se seleccionan todos los verbos que aparezcan con una frecuencia superior a 4. De este listado se selecciona un máximo de 19 términos, pues están acordes al criterio de selección especificado (tabla 26). Se decide seleccionar un máximo de 10 verbos, con las mayores frecuencias de aparición.

Para denotar la relevancia de los verbos identificados se utiliza Worllist, utilizando como estrategia de búsqueda en los 10 artículos de referencia los siguientes elementos:

1 Detección de verbos (para evadir similitudes con otro tipo de unidades).

2 Localización de unidades por lematización (facilita la extracción de todas las formas verbales).

3 Identificación de los contextos oracionales (modo de reconocer el valor de las unidades léxicas en las oraciones).

Con el objetivo de corroborar en qué orden está la relevancia semántica de estas oraciones, se realiza una comparación de las oraciones obtenidas con las que subyacen en el apartado resumen del autor de los artículos del sub-corpus. Para ello se hace un listado con todas las oraciones de cada uno de los 50 artículos seleccionados, y se analiza cuales de las oraciones extraídas mediante WorldList (que incluye de los 18 verbos) aportan los mismos contenidos que las oraciones de los resúmenes. Esto demuestra que existen verbos cuya utilización en la redacción del artículo es un factor clave para la introducción de determinado apartado. Este análisis le sirve al sistema para determinar cómo iniciar la escritura de los diversos apartados en el resumen hecho por el agente (Tabla 28).

| Verbos | Segmento |
|---|---|
| analizar<br>atrapar<br>tener | Introducción<br>Conclusiones<br>Metodología |
| generar | Resultado |
| propagar | Resultado |
| originar | Introducción |
| estudiar | Metodología |
| contribuir | Metodología |
| interrumpir | Conclusiones |
| realizar | Metodología |
| transmitir | Metodología |
| evitar | Introducción |
| Refleja | Resultado |
| conocer | Metodología |
| evaluar | Introducción |
| determinar | Introducción |
| Incluir | Conclusiones |
| delimitar | Metodología |

Tabla 28. Posición de cada verbo en los segmentos del texto

Los artículos de Ingeniería de Puertos y costas poseen altas cantidades de datos estadísticos que deben utilizarse de cara al resumen automático, sobre todo, en los apartados resultados y conclusiones, estos datos poseen alto nivel de relevancia para el resumen. Ej.

*Aproximadamente un 80% corresponde a los riegos, que son atendidos en un 82% con recursos superficiales y el resto a través de explotación de acuíferos. Las anteriores cifras explican claramente la necesidad de aumentar los recursos regulados. (76, ap.1.)*

En la mayoría de los artículos el título ha servido como estrategia de referencia a los contenidos del texto, es decir como indicador de contexto y tema esencial, aspecto importante por su relevancia para el resumen. En la tabla se muestra la relevancia de los títulos y la eliminación de las palabras vacía en ellos, lo que indica que dichos términos vacíos son irrelevantes para el resumen (ver tabla 29).

| No. | Título | Stop Word elimination |
|---|---|---|
| 1 | Criterios para el anclaje de pantalanes flotantes mediante trenes de fondeo | para, el de, de |
| 2 | Experiencia de dragado en la barra del puerto de Avilés. relación entre las características del oleaje y el crecimiento de dicha barra | de, en la, del, las, del, el y el, de |
| 3 | Propagación de ondas de borde a través de un sistema de espigones | de, de, a, de, un, de |
| 4 | Evolución de la probabilidad de fallo de diques en talud con tormentas de distintas características | de, la, de, en, con, de |
| 5 | Impacto del cambio climático en los proyectos de ingeniería de costas | del, en, los, de |
| 6 | Restauración ambiental de las marismas del joyel en la reserva natural de las marismas de Santoña y Rioja. (Cantabria) | de, las, del, en, la, de, las, de, y. |
| 7 | Variabilidad estacional del perfil de playa | del, de |
| 8 | Seguimiento del nuevo dique de San Felipe : efectos conseguidos sobre la presentación de los fenómenos de la mar de leva | del, de, la, de, los, de, la, de. |

| | | |
|---|---|---|
| 9 | Aplicación de modelos hidrodinámicos a la restauración de marismas: el caso de las marismas de Argoños y Escalante, Santoña | de, a, la, de, el, de, las, de, y. |
| 10 | Probabilidad de rotura del oleaje en condiciones de temporal | de, del, en, de |

Tabla 29. Posición de cada verbo en los segmentos del texto

En los artículos los autores declaran también topónimos y nombres geográficos que luego son elementos indispensables en sus resúmenes y en sus títulos, (ver figura) ej: Avilés, Santaña, Rioja, Cantabria, Escalante, son unidades nominales, relativas a nombres geográficos. También los autores utilizan en sus resúmenes números y cifras para especificar el estado de la investigación que están describiendo, esto tiene que ver con el tipo de artículo que se estudia, cuando el artículo tiene como género discursivo la divulgación, es decir es un artículo divulgativo estos aspectos no aparecen. Un ejemplo del uso de fórmulas se muestra a continuación ej.

*La planta piloto procesa 3 metros cúbicos por día de aguas residuales y compara el comportamiento de tres tipos de reactores anaerobios (manto de lodos, reactor empacado de flujo ascendente y descendente y reactor de lecho fluidificado o fluidizado). Los efluentes pretratados en el sistema anaerobio se someten a un sistema aerobio de discos rotatorios para convertir la materia orgánica en biomasa microbiana rica en proteína (de 15-25% en base seca, medida como nitrógeno Kjeldahl con un factor de 6.25). Los resultados obtenidos indican que el sistema, si se aplicara a tratar todas las aguas residuales generadas en las alcoholeras del país, puede producir metano con una energía equivalente aproximadamente 15 mil metros cúbicos de combustóleo al año y de 85 mil toneladas de biomasa húmeda al año (con una humedad entre el 80 y 90%) (77 resumen).*

Los autores en su mayoría realizan el resumen utilizando como formas verbales en voz pasiva y en tono impersonal, sin embargo existen, algunos trabajos donde la primera persona, aparece muchas veces en el mismo resumen del autor y en el texto.

Es importante destacar que en los análisis de los autores se ha detectado determinadas posiciones para desambiguar el texto que también serán utilizadas como regularidades para la construcción del sistema de desambiguación. Todos los resultados de este análisis se exponen en el (Anexo 17).

### 4.6.2.-Análisis de unidades léxicas que indican irrelevancia

Hasta aquí se ha observado cómo las unidades léxicas de los artículos seleccionados representan valores textuales, estilísticos y retóricos que tienen valor para un futuro resumen automático (Anexo 17). En esta investigación se hace necesaria la detección de aquellas unidades léxicas que indican irrelevancia.

Existen unidades léxicas que no son incluidas en el texto para resumen, aunque su relevancia se da de acuerdo al tipo de resumen que se quiera obtener, tal es el caso de datos estadísticos, programas computacionales, de análisis, etc. Estos son elementos que indican las técnicas empleadas para la recogida de datos en laboratorios, datos estadísticos para diversas muestras poblacionales, balances moleculares, software para la recogida de datos, etc. Ej.:

*Utilizando espectrometría de gases se obtienen los géneros fúngicos (79, ap.1)*

En el análisis de los artículos puede verse el uso profuso de tablas, gráficos e imágenes, que si bien no son utilizados para el resumen debido a que las normas para la construcción de artículos de revista no permite su inclusión en la redacción textual, sirven como una herramienta para la complementación de resúmenes no informativos, es decir sirven para la construcción de resúmenes de más alta connotación. Ej.:

*Diagrama de la planta piloto de tratamiento de vinazas ubicada en el ingenio Alianza Popular. En este marco se han hecho experimentos a lo largo de estos años que han sido ya publicados en diferentes foros (Durán de Bazúa et al, 1988; 1990; 1991; Zámano-Pérez et al, 1991). Ellos incluyen la opción de operar todos los sistemas en paralelo y una primera corrida, realizada en la zafra de 1989-90 manteniendo los sistemas anaerobios en paralelo y*

*conectando todos ellos en serie con el reactor aerobio (94, ap.3.)*

Se observa también en este estudio, que la mayoría de las oraciones refieren confirmación de experimentos y destacan hechos descritos como antecedentes de problemas, tanto de nivel ambiental como de nivel ingenieril, esto demuestra la existencia en los textos de elementos léxico-semánticos que expresan hechos y acciones irrelevantes, para el resumen informativo que es el que se pretende desarrollar en esta investigación.

Otra observación resultante del estudio de los corpus ha sido la poca presencia de datos históricos en los artículos de Ingeniería de Puertos y Costas. Los datos históricos que ocasionalmente se describen en los artículos son la fecha en que ocurren los hechos y cuando se hacen los experimentos, elementos que no se utilizan en el tipo de resumen que pretende desarrollarse en esta investigación, pero si en otras formas de representación de texto cuyo nivel de textualidad y de composición es más compleja. Ej:

En 1981 HUDC estableció un área piloto de estudio en un complejo habitacional en Akishima Tsutsujigoaka Heights, bajo el asesoramiento técnico del laboratorio del autor con el propósito de aclarar los problemas en la aplicación práctica de instalaciones de infiltración. Akishima Tsutsujigoaka Heights (de aquí en adelante referida como ATH), está localizada en los suburbios del oeste de Tokio. Un área de estudio comparativo con un sistema de drenaje convencional, se proveyó adyacente al área piloto (Fig. 1). Las áreas piloto y comparativas son de 1.61 y 1.56 has respectivamente (75, ap.1).

Como colofón de este segmento de la investigación el autor ha creído necesario presentar aquellos elementos que a partir del análisis del discurso de la temática Ingeniería de Puertos y Costas sería necesario eliminar en el resumen que se propone construir:

1. Datos estadísticos o computacionales
2. Análisis de Tablas y/o figuras.
3. Definiciones dudosas y frases de escasa consistencia científica.
4. Trabajos previos o relacionados.

## 5.7.- Análisis discursivo y sintáctico-comunicativo en Ingeniería de Puertos y Costas

Este parágrafo de la investigación se detiene en la exposición de aquellos elementos sintáctico- comunicativos del los textos de Ingeniería de Puertos y Costas para determinar reglas de inferencias a partir de la estructuración discursiva. Para ello se analizan los textos del subcorpus de referencia. Este estudio textual busca la identificación de regularidades sintáctico comunicativas en los textos anteriormente referidos para que puedan convertirse en reglas sintáctico- discursivas que faciliten la implementación computacional de regularidades textuales. Como se declaró en el capítulo 3 exponemos el análisis discursivo y sintáctico-comunicativo que realizamos de los textos del subcorpus de referencia.

Para desarrollar reglas textuales se ha apelado al perfeccionamiento de instrumentos de análisis previamente aprobados y delimitados por expertos en lingüística como Mann y Thompson (Mann y Thopson, 1988), Chomsky (Chomsky, 1965) y Moreiro (Moreiro, 2006).

En la construcción de reglas textuales se tiene en cuenta qué estructura textual caracteriza a los resúmenes, que aparecen en las publicaciones que se han utilizado para este estudio y se identifica cuál es la posición del artículo donde aparecen las oraciones con mayor carga semántica, para ser incluidas en el resumen.

Para realizar este apartado el autor se ha regido por los supuestos teóricos declarados en el capítulo 3, donde se apela a TST de Mel'cuk (Mel'cuk, 2003, Mel'cuk, 1988) y la RST de Mann y Thompson (1988), que toman como base los postulados de Beekman , Callow y Grimes (1975) y Longacre (1983) citados por Hoey (Hoey, 1983).

El procedimiento metodológico que se desarrollará en el análisis será el que se describe a continuación:

1 Analizar las estructuras discursivas (en forma de árboles formados por relaciones de Elaboración, Unión, Condición, etc.), las estructuras sintácticas (en forma de relaciones actanciales, coordinativas, atributivas

y apenditivas) y las estructuras comunicativas (en términos de Tema y Rema) existentes en los artículos del subcorpus de referencia.

2 Comprobar si en esas informaciones se dan regularidades en la estructura discursiva, sintáctica y/o comunicativa.

Se ha decido mostrar una tabla con las relaciones de discurso para que queden debidamente contextualizadas (Ver tabla 30).

| Tipo de Relación | Ejemplo |
|---|---|
| **Contraste (M)** | [Las emisiones de sulfuro en la bahía de Chichincha produjo la muerte de los peces,]N [mientras que las de la de la Bahía de Guadalupe produjeron falta de oxígeno]N |
| **Unión (M)** | [En todos los procesos de hidrólisis se realizó el un mismo proceder para estudiar las influencias de la calidad del agua en la vida de las Bahías]N y [ se terminaron los procesos al determinar que los experimentos no eran suficientes]N. |
| **Lista (M)** | [El 26 % de los fosfatos estaba mezclado con azufre]N, [el 30 % no reaccionó y el resto estaba contaminado]N |
| **Secuencia (M)** | [Se han estudiado las medidas de los mareógrafos]N, [el impacto de la variación del nivel medio del mar en los proyectos de regeneración de playas y de diques rompeolas]N. |
| **Backgroug (N-S)** | [La mayoría de los procesos de contaminación por azufre ocasiona la muerte de los peces]N. [Se ha estudiado el caso en los peces de la Bahía de Chichincha]S. |
| **Concesión (N-S)** | [Cuando se aplicó la profilaxis el nivel de oxigeno fue superior en las Bahía donde se aplicó,]N [aunque estadísticamente es un dato no crucial]S. |
| **Condición (N-S)** | [Si se aplican sustancias de bajo accionar sobre el subsuelo de la Bahía entonces es probable que]S [vuelva la vida a la fauna marina de este lugar]N. |
| **Elaboración (N-S)** | [El estudio de alternativas plantea la retirada de rellenos a la marisma y distintas posibilidades de apertura de diques]N. [Esto hace necesario un análisis de los flujos hidrodinámicos para cada alternativa propuesta, por encontrarnos frente a un estuario con una influencia mareal, donde se puede establecer un modelo que nos simule el intercambio de aguas del humedal con las carreras cíclicas mareales]S. |
| **Justificación (N-S) y** | [Se realizó un experimento para detectar los daños causados por el |

| | |
|---|---|
| **(M)** | sulfuro en la costa de Marbella]N, [por observarse la muerte de peces en 1989]S [y por la ausencia de oxigeno a varios niveles de profundidad en sus aguas]S. |
| **Propósito (N-S)** | [Para que el sulfuro pueda cumplir su función devastadora hacen falta un grupo de componentes]N [que de seguro ya estaban activados en la bahía]S |
| **Reformulación (N-S)** | No se apresa en los textos estudiados |
| **Resultado (N-S)** | [Se practicó la electrolisis en las muestras de agua]N y[ los resultados obtenidos evidencian la presencia de elevados niveles de contaminación]S |
| **Resumen (N-S)** | En la investigación se realizó la electrólisis del agua. La purificación de las sustancias ácidas] N. [En resumen se aplicaron las mejores técnicas para la investigación de residuales.]S |
| **Evidencia (N-S)** | [Se han presentado todos los datos cobre la contaminación de las bahías]N. En todos se observan daños a la fauna de nivel elevado.]S |
| **Interpretación (N-S)** | [El uso de sulfuros en la eliminación de mosquitos es perjudicial en los ríos es perjudicial]N, [por tanto no es pertinente su implementación como mecanismo profiláctico]S. |

Tabla 30. Posición de cada verbo en los segmentos del texto

Estas son las relaciones retóricas que parten de la teoría de Mann y Thompson (Mann and Thompson, 1990). A modo de ejemplo se ha decidido mostrar un artículo científico de nuestro subcorpus de referencia para ilustrar en el contexto aquellos elementos discursivos que se han declarado (Ver tabla 31).

| Artículo del Subcorpus |
|---|
| **Introducción**<br>[A través de los tiempos las marismas de Santoña han experimentado una importante reducción de su área inundable]N. [*Esta reducción ha estado motivada, en parte, por la naturaleza sedimentaria propia de este tipo de estuarios y, en mayor medida, por la intervención humana]* S. *[Un ejemplo de este último caso, es la afección acontecida por la construcción de la carretera Cicero- Santoña, conocida como la "Carretera de los Puentes"] S.*<br>***Propósito***<br>[En los últimos años, el reconocimiento de la enorme importancia que estas zonas inundables tienen tanto por su gran productividad biológica como por su capacidad de laminación de avenidas y su valor paisajístico]N, *[ha dado lugar a una política proteccionista y regeneradora de estos ámbitos estuarinos]*S. ***Circunstancia*** |

[Dentro de esta política se enmarcó el estudio que se presenta, en el que, con motivo de las obras de mejora de la "Carretera de los Puentes" se pretendió: (1) recuperar ambientalmente un área de estuario que fue alterada cuando se construyó la citada carretera, ]N [ (2) minimizar los efectos negativos que pudieran ocurrir durante la construcción de dos puentes en la marisma, analizando los efectos que cierres temporales y parciales de los actuales canales de desagüe pudieran tener en los procesos de inundación y de erosión/sedimentación de la marisma,]N [y (3) calibrar el modelo hidrodinámico con medidas efectuadas durante la ejecución de las obras.]N

***Justificación***

**2. Metodología**

[Cuando en un canal del estuario se produce un estrechamiento de sección se originan unas pérdidas de carga localizadas] que conllevan: (a) variación del régimen de inundación del estuario] S, [(b) desfases de la onda de marea y (c)] S [variación del régimen de velocidades.]S

***Circunstancia***

[Para minimizar estos efectos, producidos por los estrechamientos realizados durante las obras en la marisma]S, [surgió la necesidad de establecer unos criterios que permitieran decidir si la afección era o no admisible] N ***Circunstancia***. [Los criterios que se adoptaron en el estudio fueron diferentes para cada puente]N, [debido a que en el caso del puente de Argoños existían otros dos canales que comunicaban la marisma de Argoños con el resto del estuario, circunstancia que originaba un funcionamiento distinto del que se daba en el puente de Escalante, al ser este último la única sección por la que discurría la ría del mismo nombre]***Circunstancia***

[La simulación numérica de las distintas situaciones de estudio se realizaron con el modelo hidrodinámico H2D desarrollado por el Grupo de Ingeniería Oceanográfica y de Costas de la Universidad de Cantabria]N. [El modelo resuelve las ecuaciones de onda larga promediadas en vertical, mediante un algoritmo en diferencias finitas implícito de doble barrido]N ***Unión***

**3. Resultados**

[Como resultado del estudio se obtuvieron curvas de diseño para ambos puentes]N [que a su vez se utilizaron para determinar los días que se podían cerrar los puentes un determinado tanto por ciento sin incumplir los criterios adoptados] S ***Resultado*** (figura 2). [En el artículo final se expondrán los criterios]N [que se establecieron para la minimización de los efectos ambientales, así como la metodología de uso de los modelos hidrodinámicos para alcanzar los objetivos de no afección.] S ***Resumen***

Tabla 31. Ejemplo de relaciones de cohesión en el texto

Estas unidades que han sido señaladas en el texto se utilizan en los resúmenes realizados por los autores en las revistas, pues son unidades que comúnmente son descritas por los autores en el momento de escribir el

resumen de sus artículos (Anexo 12).

### 5.7.1- Desarrollo de las Reglas sintáctico-comunicativas del resumen

De acuerdo con las coincidencias registradas mediante el estudio de los 50 artículos de publicaciones del ámbito de la Ingeniería de Puertos y Costas se especifica el proceder que debe ser la guía para el desarrollo de la aplicación. Al igual que en los trabajos de D´Cunha, (2006) se optó por dividir las unidades regulares en tres niveles de procesamiento: textual, léxico y discursivo-sintáctico-comunicativo, para cada uno de estos niveles se desarrollan reglas específicas.

La forma de operacionalizar el resumen se basa específicamente en lo que se plantea a continuación:

1. Supresión de oraciones, sentencias o fracciones textuales, asumiendo criterios basados en unidades léxicas.
2. Eliminar oraciones tomando como referente las regularidades discursivas y sintácticas comunicativas del texto.

### 4.7.2.2-Desarrollo de reglas discursivo-sintáctico-comunicativas

Según los estudios de Álvarez de Mon y Rego (Alvarez de Mon, 1999) los textos científicos poseen similitudes expresivas, sintácticas, cohesivas y enunciativas que facilitan su similitud dada su forma de construcción, por ello estas reglas que se declaran toman el enfoque de D´Cunha (2006). (Anexo 12).

Como se declaró en el modelo (capítulo 4) a partir de las regularidades detectadas en los textos es posible realizar un estudio minucioso de la estructura sintáctico-comunicativa del artículo. Para realizar este análisis se contrastan los textos contra los resúmenes de los artículos especializados en Ingeniería de Puertos y Costas.

Es importante declarar que se ha declarado como regularidad lo siguiente:

1) Encontrar al menos tres casos en los que exista coincidencia.

2) Constatar mediante el análisis del artículo las presencia de determinada regularidad en el resumen.

**A) Regularidades en las que se elimina un satélite discursivo**

Es indiscutible la presencia de nexos comunicativos en forma Núcleo- Satelital, cuestiones que los autores de los resúmenes eliminan cuando construyen un sumario (Tabla 32-37). Los satélites que se eliminan están en correspondencia con las relaciones discursivas:

a) Concesión.

b) Reformulación.

c) Resultado.

d) Justificación.

e) Circunstancia.

f) Propósito.

El autor ha decido proponer algunos ejemplos de cada una de las relaciones discursivas de acuerdo a los resumes de los autores.

a) Casos de eliminación del satélite de Concesión

| Fragmento del artículo 89 |
| --- |
| [En este estudio se declaran los problemas toxicológicos producidos por el exceso de emisiones de sulfuro a la bahía de Chichincha]N [Se ha dicho en muchas ocasiones que los referidos estudios siempre son iguales.]S |
| **Fragmento del Resumen del autor** |
| *Se desarrolla un estudio sobre las afectaciones toxicológicas que ha producido el sulfuro en la Bahía de Chichincha* |

Tabla 32. Eliminación de un satélite de concesión.

b) Casos de eliminación del satélite de Reformulación

| Fragmento del artículo 209 |
| --- |
| [La grandes dificultades que presentan los peces de las costas de Granada, son las emisiones de desechos sólidos]N [ esto obliga a desarrollar planes de desarrollo específico para esta franja costera]S<br>**Fragmento del Resumen del autor**<br>*Se realiza un estudio de los peces de la costa granadina por los efectos de las emisiones de desechos sólidos* |

Tabla 33. Eliminación de un satélite de Reformulación

c) Casos de eliminación del satélite de Resultado

| Fragmento del artículo 305 |
| --- |
| [Los resultados del estudio muestran dos vertientes: una nos muestra las dificultades generadas por la baja calidad del agua y la otra evidencia el poco nivel de conciencia de la población con respecto a este problema]N [ se entienden estas dificultades como los problemas de oxigeno en los peces de esta franja costera]S<br>**Fragmento del Resumen del autor**<br>*Se realiza un estudio que evidencia dificultades con la calidad del agua y con las concientización de la población.* |

Tabla 34. Eliminación de un satélite de Resultado

d) Casos de eliminación del satélite de Justificación

| Fragmento del artículo 405 |
| --- |
| [Debido a la inocencia de vertimientos de aguas tóxicas en el CAI “Antonio Finalet ]N [pues ha creado deficiencia en el desarrollo de fitoplancton en las orilla de la costa]S<br>**Fragmento del Resumen del autor**<br>*Se pretende estudiar la incidencia de los vertimientos de las aguas tóxicas del CAI “Antonio Finalet.* |

Tabla 35. Eliminación de un satélite de Justificación

e) Casos de eliminación del satélite de Propósito

| Fragmento del artículo 206 |
| --- |
| [Para realizar esta investigación se desarrollaron estudios de espectrometría de gases en las aguas de la bahía]N [El desarrollo de estos estudios son elementalmente descritos en los trabajos Hubberk]S<br>**Fragmento del Resumen del autor**<br>*Se presenta un estudio espectrométrico utilizando gases inertes para medir la conservación de especies.* |

Tabla 36. Eliminación de un satélite de Propósito

f) Casos de eliminación del satélite de Circunstancia

| Fragmento del artículo 206 |
| --- |
| [En el año 2009 en la bahía de la Habana se desarrolló una investigación para estudiar la muerte del fitoplancton]N [Dicha investigación se encargaba esencialmente de analizar los efectos de la contaminación sobre la vida animal]S<br>**Fragmento del Resumen del autor**<br>*Las condiciones de la bahía de la Habana son desfavorables para el desarrollo de la vida del fitoplancton.* |

Tabla 37. Eliminación de un satélite de Circunstancia

**B) Regularidades en las que se elimina un núcleo discursivo**

En nuestro análisis de los textos hemos detectado la eliminación de los núcleos de evidencia y de interpretación de forma Núcleo- Satélite (Tabla 36 - 37).

*Interpretación*

a) Casos de eliminación del núcleo de Interpretación

| Fragmento del artículo 138 |
| --- |
| [ Las cifras de contaminación son elevadas para esta zona portuaria por lo que es necesario eliminar los vertimientos, pues el 18 % de su fauna puede ser afectada]N [Los datos revelan la factibilidad de estudios de contaminación en la zona]S<br>**Fragmento del Resumen del autor**<br>*La contaminación en las zonas portuarias se debe esencialmente la existencia de vertimientos,* |

| |
|---|
| *que producen que el 18 % de la fauna sea afectada. .* |

Tabla 36. Eliminación de un núcleo de Interpretación

### b) Casos de eliminación del núcleo de Evidencia

| Fragmento del artículo 11 |
|---|
| [La fauna de la zona del canal está en peligro de extinción debido al dragado, estudios recientes declaran que el 23 % de dicha fauna está desapareciendo]N [Esta situación obliga a dar parte a las autoridades sobre la situación para que se proceda a la intervención de la zona]S<br>**Fragmento del Resumen del autor**<br>*La fauna de la zona está en peligro de extinción debido a los procesos de dragado de la bahía, es por ello que los expertos han hecho reportes donde muestran datos sobre la posible extinción de las especies.* |

Tabla 37. Eliminación de un núcleo de Evidencia

### C) Regularidades en las que no se separa el satélite de su núcleo

En muchas ocasiones se ha observado que los autores, no eliminan algunas relaciones discursivas Núcleo – Satélite debido a que se segmenta la información y se pierde información relevante para los lectores. Esta situación se da esencialmente cuando estamos frente a relaciones de Condición y de Resumen con respecto a sus satélites (Tabla 38 – 39).

### a) Casos de mantenimiento del satélite de Condición.

| Fragmento del artículo 17 |
|---|
| [El estudio se realiza bajo la supervisión de expertos de la COMARNA]N [Estos supervisores tiene la obligación absoluta y especifica se controlar la calidad del experimento ]S [Los procesos de calidad en estos experimentos son elevado coste para las empresas]N[La calidad es esencial en estos estudios para que sean confiables]S<br>**Fragmento del Resumen del autor**<br>*La supervisión de los expertos es necesaria para el desarrollo de experimentos de toxicidad, aunque el coste para las empresas sea elevado...* |

Tabla 38. Mantenimiento del Satélite de Condición.

**b) Casos de mantenimiento del satélite de Resumen.**

| Fragmento del artículo 48 |
|---|
| [La escasez de agua es común en muchas áreas, principalmente en el Oeste Interior, aún en el sur de Ontario, una península construida en el corazón del cuerpo de agua dulce más grande del mundo.]N S [Los canadienses se han dado cuenta lentamente de que el 9% del suministro anual de agua renovable en el mundo no es, por sí sólo, una garantía contra la escasez.<br>**Fragmento del Resumen del autor**<br>*La escasez de agua en el lago Notario ha hecho que los canadienses tomen conciencia sobre el peligro de la pérdida de las aguas de la tierra...* |

Tabla 39. Mantenimiento del Satélite de Resumen

**D) Regularidades en las que no se separan dos núcleos**

Siguiendo la línea de regularidades se ha observado como los autores que en relaciones multinucleares de tipo: Lista, Contraste y Unión (Tabla 40 – 42).

a) No se separa la Lista y sus núcleos

| Fragmento del artículo 19 |
|---|
| Para realizar esta investigación clasificó el agua en grupos: [1) agua con alto nivel de acidez (60 militros),]N [, 2) agua con bajo nivel de hidrógeno (20 mililitros)]N [y 3 agua con bajo nivel de oxigenación]<br>**Fragmento del Resumen del autor**<br>*Se establecen 3 clasificaciones para las aguas: la primera incluye agua ácida (60 militros), la segunda agua escasa de hidrógeno (20 militros) y el tercero agua con bajo nivel de oxigenación.* |

Tabla 40. No supresión de los núcleos en la relación de lista

c) No se eliminan los Núcleos asociados a una relación de Unión

| Fragmento del artículo 92 |
|---|
| [El objeto de este estudio es conocer la incidencia de los elevados niveles de salinidad en las costas de Cantabria en el período 2004-2005,]N [y la existencia de de medios para la contención de estos problemas debido a la experiencia internacional]N<br>**Fragmento del Resumen del autor** |

| |
|---|
| *Dada la experiencia internacional del uso de medios de contención es posible estudiar la contención de la salinidad en la costa cantábrica.* |

Tabla 41. Mantenimiento de los núcleos asociados a una relación de unión

d) En presencia de relación multinuclear de Contraste no se eliminan sus núcleos.

| Fragmento del artículo 101 |
|---|
| [Las concentraciones de ácido sulfúrico en el SC fueron superiores a las halladas en la costa de Siguanea (345,9 [244,4] ng/ml en comparación con 272,6 [299,3] obtenido en Cancún con el error de probabilidad de un (0,05).]N [ En el resto de los elementos de evaluación los valores son estables ]N |
| **Fragmento del Resumen del autor** |
| *[Las concentraciones de ácido sulfúrico en el SC fueron superiores a las halladas en la costa de Siguanea (345,9 [244,4] ng/ml en comparación con 272,6 [299,3] obtenido en Cancún con el error de probabilidad de un (0,05).]N [ En el resto de los elementos de evaluación los valores son estables ]N* |

Tabla 42. Mantenimiento de los núcleos en la relación de contraste

**F) Regularidades en las que se elimina un satélite discursivo relacionado con un elemento sintáctico.**

El autor en sus observaciones también localizó algunas cuestiones que expresan detalles lingüísticos con nexos sintácticos, regularidades que relacionan estructuras lingüísticas, sintácticas y discursivas. La coincidencia de elementos de Elaboración (satélite discursivo) y la posición de determinados atributos sintáctico explicativos, no son incluidos en los sumarios por los autores. Esa cuestión se da con la relación (APPEND) (Tabla 43).

| Fragmento del artículo 103 |
|---|
| El nivel de investigación se ha elevado considerablemente en lo referente a la espectrometría de gases, esto hace que el análisis con esta técnica ya no sea tan demorado en dependencia de la muestra que se selección, sea muy eficiente, por lo que se ha decidido declarar el estudio buscando gases inertes poco pesados. |
| **Fragmento del Resumen del autor** |
| *El nivel de investigación se ha elevado considerablemente en lo referente a la espectrometría de gases, esto hace que el análisis con esta técnica, sea muy eficiente, por lo que se ha decidido declarar el estudio buscando gases inertes poco pesados* |

Tabla 43. Regularidades en que se elimina un satélite discursivo

**G) Regularidades en las que se elimina un satélite discursivo relacionado con un elemento comunicativo**

Otras estructuras pueden también relacionarse con la estructura comunicativa y discursiva de los artículos, especialmente si existen nexos entre el Tema, los autores excluyen estos elementos de sus construcciones textuales (Tabla 44).

| Fragmento del artículo 99 |
|---|
| [Como grupo control se empleó uno constituido por 456 personas residentes en la bahía.]N [Este grupo se obtuvo mediante la selección aleatoria de las personas que viven en el lugar]S<br>**Fragmento del Resumen del autor**<br>*[Como grupo control se empleó uno constituido por 456 personas residentes en la bahía.]N [Este grupo se obtuvo mediante la selección aleatoria de las personas que viven en el lugar]S* |

Tabla 44. Regularidades en que se elimina un satélite discursivo con nexos con un elemento comunicativo

**H) Regularidades en las que no se elimina un satélite discursivo relacionado con un elemento comunicativo**

Se ha observado que si el contenido de un satélite discursivo de Elaboración se refiere al Rema de su núcleo, los expertos deciden eliminarlos de su núcleo (Tabla 45).

| Fragmento del artículo 99 |
|---|
| [El objeto de esta investigación es determinar el porcentaje de áreas de alto nivel de acidez en la cuenca del río Máximo, durante el año 2004.]N [ Este análisis no ha sido realizado en Cuba ]S<br>**Fragmento del Resumen del autor**<br>*El estudio de los niveles de acidez de los ríos cercanos a los CAI no ha sido efectuado ni divulgado en Cuba...* |

Tabla 45. Regularidades en que no se elimina un satélite discursivo con nexos con un elemento comunicativo

### *5.8.- Formalización de Reglas para la Extracción del Texto*

Como ya se ha expuesto en el capítulo tres de la investigación se han

desarrollado estudios lingüísticos y estructurales para formalizar reglas de acuerdo a la tipología de sumario que se pretende construir, las mismas, están declaradas en consonancia con los análisis de diversos apartados de los artículos de Ingeniería de Puertos y Costas. Como son reglas estas responden a regularidades desde el punto de vista léxico-semántico, estructural y de nivel de cohesión, al igual que en los trabajos de D´cunha (2006) y Paneca (Paneca, 2009) se han clasificado las normas de la siguiente forma atendiendo a sus rasgos esenciales:

1 Estructura textual.

2 Unidades léxicas

3 Estructura discursiva y sintáctico-comunicativa.

Desde el punto de vista de implementación, se ha intentado hacer que las reglas formalizadas coincidan con los procesos de implementación del resumen, con las operatorias y la estrategia de construcción del extracto. Estas reglas se unen a las posturas de lectura y son implementadas por agentes de software cuyo nivel de especialización en la lectura y construcción del texto ofrece confiabilidad en la calidad del proceso. De esta forma el agente de lectura y el de resumen asumen diversos niveles textuales para construir un extracto:

1. Nivel textual: consiste en la implementación de reglas basadas en la estructura del texto (denominadas reglas textuales).
2. Nivel léxico: Detección, eliminación y selección de unidades léxicas de la base de conocimientos y de la ontología del sistema.
3. Nivel discursivo y sintáctico-comunicativo: Formalización de normativas de las reglas discursivo-sintáctico-comunicativas (reglas PUERTOTERM).

El análisis de los corpus en este capítulo nos ha mostrado la existencia de una estructura retórica (IOMRC). A igual que los estudios de María Pinto (Pinto, 2004, Pinto, 2001) y Hernández (Hernández, 2006). Esto da pie a la existencia de una solución que debe funcionar a nivel de regla, realizada por el agente de

lectura declarado en la metodología. Esta regla que vamos a llamar modelo de reconocimiento de texto será encargada de localizar en el texto las unidades retóricas específicas el texto, es decir los apartados (Introducción, Metodología, Resultados y Conclusiones). Dicha regla se construye debido a que los autores utilizan en los resúmenes elementos de los apartados anteriormente definidos.

Partiendo de los criterios de Lunh (Lunh, 1958) y las modificaciones de Mathis, (Mathis et al., 1973) se ha constatado la posición estratégica de determinadas oraciones en los segmentos de los textos analizados, las cuales evidencian mayor relevancia que otras en el momento de la selección del resumen (estas posiciones ya han sido declaradas en el parágrafo 5.5.2.3).

A modo de táctica operacional para este modelo de resumen se siguen los mismos supuestos teóricos de Lunh (1958) y Salton (Salton, 1996) donde los modelos de resumen utilizan una aproximación por extracción donde ocurren procesos de transformación de corpus en los cuales se convierte el texto en tokens y se eliminan los signos de puntuación existentes en su volumen. Las acciones propuestas para el modelo TEXMINER son las siguientes:

Eliminar oraciones o fragmentos del texto.

a) Asignar un peso ponderado a cada oración del texto según la presencia en ella de vocablos y su función o posición en el texto.

b) Suprimir oraciones de acuerdo al poco peso en sus niveles semánticos.

A partir de esta forma ya clásica en el tratamiento del resumen se ha decidido desarrollar diversos tipos de reglas para operar en texto científico:

I. Reglas que suprimen secuencias del texto fuente, oraciones y/o pasajes textuales con determinado nivel léxico-semántico.

II. Reglas que delimitan aquellas oraciones que en el texto original han de ser eliminadas y almacenadas para otros tipos de construcción

III. Reglas que obligan a buscar en otros niveles de descripción si los candidatos oracionales no son correctos, es decir buscar en la descripción de otros niveles de texto (imágenes y gráficos).

IV. Reglas que aportan relevancia a las oraciones en dependencia del lugar donde se encuentren y a la presencia en ellas de vocablos o términos de las listas de prioridad.

V. Reglas que permiten desarrollar estrategias de desambiguación léxica.

De esta manera (Ver Figura 37), el modelo de extracción de texto asume dos fases con las siguientes particularidades:

Fase I:

1 Constituye la supresión de oraciones y pasajes textuales de bajo nivel de relevancia en el texto (usar reglas del tipo I).

Fase II:

1 Se proponen oraciones o fragmentos de texto candidatos a ser eliminados (mediante las reglas del tipo II).

2 Si no existen oraciones que cumplan este criterio buscar otros textos asociados al texto que permitan mejorar la descripción. Reglas de Tipo III.

3 Descarte aquellas oraciones que poseen menos nivel de puntuación y facilite la proyección de aquellas que poseen más alto nivel de relevancia.

4 Aplique reglas de sintaxis y comunicación para lograr la cohesión del texto.

5 Desambigüe el texto para brindar mayor nivel de claridad en la lectura de la información. (reglas IV).

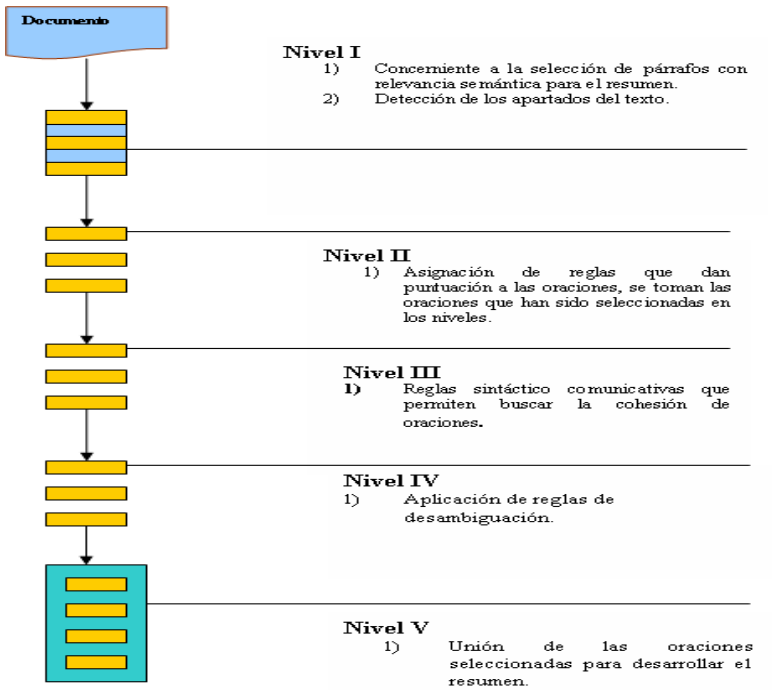



Figura 37. Modelo de Procesamiento del Texto elaborado a partir de la concepción de las reglas

La aplicación exacta de estos elementos en el sistema de desambiguación y extracción quedan de la siguiente forma.

6 Nivel 1: Es utilizado y visualizado como un nivel de segmentación de texto, donde al igual que en varios estudios como los de Enders-Nigemeyer (2005) se delimita el texto en 4 apartados donde los contenidos poseen mayor nivel de preponderancia. En ocasiones estos apartados no están delimitados, esto obliga a que el agente de lectura reconozca los elementos por los que tradicionalmente se marcan estos segmentos del texto y confronte la base de sinónimos donde aparecen sinónimos de estos apartados, al identificar los sinónimos le asignará la estructura seleccionada para este tipo de estudio (IOMRC). Es importante destacar que otras reglas para el tratamiento del texto se aplicarán en otros niveles de desarrollo del texto en la Fase II, donde se construye un resumen basado en extracción-abstracción.

7 Un nivel léxico (2), utiliza normas basadas en unidades léxicas que facilitan la asignación de relevancias a determinados segmentos

textuales u/o oraciones. Las oraciones al ser eliminadas pueden ser confrontadas con otros textos donde la descripción sea mejor desde el punto de vista léxico-semántico. Estos elementos son claramente explicados dentro de este capítulo de la investigación. En este nivel, se aplican reglas basadas en unidades léxicas que asignan puntuación a oraciones.

8 Un nivel discursivo y sintáctico-comunicativo (3), consistente en la aplicación de reglas (PUERTOTERM), donde, de acuerdo a cada segmento textual se aplica determinado elemento de cada relación discursiva, regularidades que son enriquecidas semánticamente mediante bases de conocimientos (esto se explicitará en el capítulo de desarrollo de la aplicación). En estas reglas dictan qué elementos discursivos se mantienen o se eliminan de acuerdo a las marcas de cohesión realizadas en el texto, estas reglas están en consonancia con la fase uno de la etapa de extracción de texto. Esto permite obtener un resumen por extracción-abstracción.

9 En el modelo de D´Cunha (2006) se tiene en cuenta la longitud de las oraciones deseadas por el usuario, sin embargo, no consta que estos criterios sean necesarios para la calidad del resumen. En el modelo extractivo que se describe el autor ha creído pertinente desarrollar una fase de desambiguación (Nivel 4) a nivel oracional utilizando mejoras en las oraciones resultantes de las dos fases I, II y III.

10 La fase final está constituida por la unión de las oraciones de todos los apartados (Nivel V).

### *5.8.1.- Criterios lingüísticos del modelo*

Después de establecer los elementos que sirven de esclarecimiento de los procesos de extracción que debe seguir el artículo científico en el dominio de la Ingeniería de Puertos y Costas, es necesario entonces, declarar aquellas reglas que facilitan la extracción y la desambiguación del texto para que puedan ser aplicadas en la etapa final.

#### 5.8.1.1.- Criterios textuales: formalización de reglas textuales

Las reglas textuales tienen como criterio esencial, el papel preponderante que ocupan determinadas oraciones en el texto y las cargas de significado específico para cada apartado del texto, además la posición de las sentencias u oraciones descritas en el texto (Figura 38 - 44). Según el análisis de los textos son estas las oraciones y los apartados de mayor carga semántica en cada texto:

1 *3 primeras y 4 últimas oraciones* (Introducción)

2 *3 primeras y 4 últimas* (Metodología)

3 *3 últimas oraciones* (Resultados)

4 *5 primeras o 3 últimas oraciones* (Conclusiones)

**Formalización de las reglas textuales**

| 1ra regla textual |
| --- |
| *Si en la oración es una de las tres primeras oraciones del apartado Introducción, entonces* $A^2:=A^2+A^s$ |

Figura 38. 1ra regla textual

| 2da regla textual |
| --- |
| *Si la oración es una de las 4 últimas oraciones del apartado Introducción, entonces* $A^2:=A^2+A^s$ |

Figura 39. 1ra regla textual

| 3ra regla textual |
| --- |
| *Si en la oración es una de las tres primeras oraciones del apartado Metodología, entonces* $A^2:=A^2+A^s$ |

Figura 40. 3ra regla textual

| 4ta regla textual |
| --- |
| *Si en la oración es una de las tres últimas oraciones del apartado Metodología, entonces* $A^2:=A^2+A^s$ |

Figura 41. 4ta regla textual

| 5ta regla textual |
| --- |
| *Si en la oración es una de las tres primeras oraciones del apartado Metodología, entonces* $A^2:=A^2+A^s$ |

Figura 42. 5ta regla textual

| 6ta regla textual |
| --- |
| *Si en la oración es una de las 5 primera oraciones del apartado Conclusiones, entonces* $A^2:=A^2+A^s$ |

Figura 43. 6ta regla textual

| 7ma regla textual |
| --- |
| *Si en la oración es una de las 4 últimas oraciones del apartado Conclusiones, entonces* $A^2:=A^2+A^s$ |

Figura 44. 7ma regla textual

### *5.8.2.-Criterios para la construcción de Reglas que asignan puntuación a las oraciones (Reglas Léxicas)*

Para asignar puntuación a oraciones, se han desarrollado reglas que indican la relevancia otorgada por los expertos a ciertos grupos textuales y oracionales, en los artículos del corpus español. Las unidades léxicas que representan al artículo en Ingeniería de Puertos y Costas. Este criterio toma como base el estudio de elementos relevantes en las oraciones (Ver Anexo 4).

La indiscutible presencia en el texto de unidades léxicas nominales como: realiza, presenta, se ha considerado, cabe destacar, descrita, fin, estudia, ,resultados, comparación, resultado, validación, construido, analizar, objetivo, objeto, propósito, método, intención, conclusión, resultado, facilita la ponderación del texto.

Teniendo en cuenta la presencia de estas unidades léxicas la ponderación será sumativa, es decir si aparecen una unidad léxica, el valor de la oración será 1, si posee dos unidades léxicas, será 2, cero será si no posee ninguna. Mientras más unidades léxicas tengan la oración, mayor valor tendrá para el resumen esa oración candidata. También se ha observado que la presencia de más de

una unidad léxica en oraciones compuestas, obliga a reducir a cero el valor de la sentencia (Figura 45 – 47).

| 1ra regla léxica |
| --- |
| *Si en la oración aparece una, dos o tres unidades léxicas entonces* $A^2:=A^2+A^s$ |

Figura 45. 1ra regla léxica

| 2da Regla Léxica |
| --- |
| *Si en la oración es compuesta y posee las unidades léxicas de la lista entonces, se suprime y la puntuación entonces es cero* $A^2:=A^2-A^s=0$ |

Figura 46. 2da Regla Léxica

La presencia en el texto de unidades relativas al título de la obra le dará a la oración un valor 2, por tanto será elegible esta oración al aumentar su peso sobre las demás.

| 3ra Regla Léxica |
| --- |
| *Si en la oración aparecen unidades relativas al título del artículo (solo palabras, no enlaces ni elementos vacíos eliminados por stop Word elimination)* $A^2:=A^2+A^s+A^s=2$ |

Figura 47. 3ra Regla Léxica

Algo que también se ha visto en estos estudios es la presencia del texto en voz pasiva, activa y en determinadas ocasiones en primera persona. Se ha decidido dejar solo la voz pasiva para normar la actividad, pues en algunas ocasiones los artículos se escriben de forma impersonal (Figura 48- 53).

| 4ta Regla Léxica |
| --- |
| *Si la oración aparece redactada en voz pasiva, es decir los si las formas verbales están todas en voz pasiva entonces* $A^2:=A^2-A^s$ |

Figura 48. Regla Léxica

| 5ta Regla Léxica |
| --- |
| *Si la oración aparece redactada en voz primera persona, es decir los si las formas verbales están todas en primera persona entonces* $A^2:=A^2+A^s$ |

Figura 49. 5ta Regla Léxica

#### 5.8.2.1- Reglas que asignan puntuación por mostrar elementos estadísticos

| 6ta Regla Léxica |
| --- |
| *Si la oraciones seleccionadas como relevantes en los apartados metodología y conclusiones aparecen datos estadísticos , asigne un punto más a la oración por relevancia (se asigna más valor si aparecen más elementos)* $A^2:=A^2+A^s+A^s=2$ |

Figura 50. Reglas que asignan puntuación por mostrar elementos estadísticos

#### 5.8.2.2.-Reglas que asignan puntuación por mostrar elementos Químicos

| 7ma Regla Léxica |
| --- |
| *Si la oraciones seleccionadas como relevantes en los apartados metodología y conclusiones aparecen fórmulas químicas, asigne un punto más a la oración por relevancia (mientras más fórmulas aparezcan mayor será el valor)* $A^2:=A^2+A^s+A^s=2$ |

Figura 51. Reglas que asignan puntuación por mostrar elementos Químicos

Reglas que asignan puntuación por la presencia de nombres geográficos

| 8va Regla Léxica |
| --- |
| *Si la oraciones seleccionadas como relevantes en los apartados introducción, metodología y conclusiones aparecen nombres geográficos entonces, asigne un punto más a la oración por relevancia(igual a los otros incisos)* $A^2:=A^2+A^s+A^s=2$ |

Figura 52. Reglas que asignan puntuación por la presencia de nombres geográficos

#### 5.8.2.4.-Reglas que declaran unidades léxicas nominales de la lista desarrollada para el dominio.

| 9na Regla Léxica |
| --- |
| *Si la oraciones seleccionadas como relevantes en los apartados introducción, metodología y conclusiones aparecen unidades nominales (verbos de la lista) entonces, asigne un punto más a la oración por relevancia (igual los restantes incisos)* $A^2:=A^2+A^s+A^s=2$ |

Figura 53. Reglas que declaran unidades léxicas nominales

**Reglas que asignan puntuación a aquellas oraciones que han sido declaradas a partir de otros textos como imágenes, tablas, gráficos, etc.**

Generalmente en los trabajos que se han estudiado se aprecia la presencia de datos que están construidos a partir del análisis de tablas y gráficos, existentes en el texto, es decir a partir de una forma diferente de textualidad. Como en el sistema que se diseñará existen también descripciones resumidas que provienen de imágenes, con las cuales se pueden construir resúmenes de mayor nivel de construcción si el usuario lo desea se declara esta regla (Figura 54).

| 9na Regla Léxica |
| --- |
| *Si la oraciones seleccionadas provienen de otra forma textual, o de un nuevo tipo de procesamiento de la información proceda de la siguiente forma (igual los restantes incisos)* $A^2:=A^2+A^s+A^s=2$ |

Figura 54. Reglas que asignan puntuación a aquellas oraciones que han sido declaradas a partir de otros textos como imágenes, tablas, gráficos, etc.

### 5.8.2.5.1.-Reglas léxicas que eliminan oraciones

La supresión de oraciones es una de las tareas que más se repite en los corpus que se han analizado, pues en los artículos hay un gran número de ellas que son eliminadas, por ser irrelevantes para la construcción de un resumen. Esta regularidad sirve de base o sustento de praxis a las reglas que proponemos a continuación. Estas reglas están declaradas a partir de las regularidades descritas en el tópico anterior (Anexo 12).

Una regularidad en el tratamiento del texto ha sido siempre la ausencia de datos que expongan informaciones relacionadas con software computacional y procesos estadísticos de selección de muestras y poblaciones. A partir de esta realidad desarrollamos reglas que facilitan la supresión de aquellos elementos se han declarado (Figura 55).

| 10ma Regla Léxica |
| --- |
| *Si una oración contiene información sobre datos estadísticos y computacionales encerrados en paréntesis. Then eliminate Parenthesis from sentence (s) EISE IF* / $A^{prso}$ */> 1* eliminate (s) from $A^{prso}$ |

Figura 55. 10ma Regla Léxica (Reglas que eliminan oraciones)

No se incluyen generalmente en el texto de resumen referencias a tablas y gráficos, aunque si su interpretación. Elimine todo dato referente a tablas y gráficos en el texto candidato para el resumen (Figura 54).

| 11na Regla Léxica |
| --- |
| *Si una oración contiene información sobre referencia a tablas y gráficos, elimine las oraciones que contienen dichos datos. Then eliminate (s) from* / $A^{prso}$ / |

Figura 56. 11na Regla Léxica (Reglas que eliminan oraciones)

No se aprecia en los trabajos analizados en este corpus la presencia de información relativa a trabajos previos. Los criterios de construcción textual en las publicaciones seriadas eliminan esta información dentro de estos apartados y en los resúmenes de los autores no se aprecian dichos datos (Figura 56).

| 12 ma Regla Léxica |
| --- |
| *Si una oración contiene información relativa a trabajos previos entonces aplique lo que se indica. Then eliminate (s) from/* $A^{prso}$ / |

Figura 57. 12 ma Regla Léxica (Reglas que eliminan oraciones)

Como última regla declaramos una regularidad que suprime aquellas oraciones que no poseen formas verbales, pues generalmente están declaradas por inicio y final por puntos. Al igual que en el trabajo de D´Cunha (2006) se intenta llevar al agente a la supresión de aquellas estructuras textuales de forma del artículo, los que son definidos por (I(O) MRC).

### *5.9.- Criterios discursivos y sintáctico-comunicativos: formalización de reglas discursivo-sintáctico-comunicativas*

Un elemento que se ha descrito en el modelo de resumen son las estructuras discursivas y sintácticas comunicativas de acuerdo a los modelos de cohesión más avanzados en materia de textos electrónicos. Como ya se ha dicho estos

análisis de regularidades se basan específicamente en los modelos de construcción desarrollados por los autores de los textos en este dominio.

Estas regularidades traen como consecuencia las reglas PUERTOTERM que se basan en las DISICO (D´cunha, 2006), a las cuales supera en casos y en estrategias de construcción. Las reglas que se desarrollan son aplicadas a los siguientes casos: Criterios de supresión de satélites discursivos, Criterios de supresión de núcleos discursivos, Criterios de supresión de satélites, Criterios de desambiguación léxica.

### 5.9.1.- Criterios de supresión de satélites discursivos

Se ha apreciado la supresión de satélites de discurso mientras aparecen expresiones de Fondo, Concesión, Reformulación, Resultado y Circunstancia (Figura 58 - 62).

**A) Suprimir el satélite discursivo en presencia de relaciones de: a) Background, b) Concesión, c) Reformulación, d) Resultado, e) Justificación y f) Circunstancia.**

a) Background (Fondo)

| 1 ra Regla PUERTOTERM |
| --- |
| *Si en una oración aparece un satélite como una relación de fondo entonces elimínela: IF (s) is satellite of a BACKGROUND relation B Entonces si /A$^{prsa}$/>1 Eliminate S* |

Figura 58. 1 ra Regla PUERTOTERM

b) Concesión

| 2 da Regla PUERTOTERM |
| --- |
| *Si en una oración aparece un satélite como una relación de fondo entonces elimínela: IF (s) is satellite of a CONCESION relation B Entonces si /A$^{prsa}$/>1 Eliminate S* |

Figura 59. 2 da Regla PUERTOTERM

c) Resultado

| 3 ra Regla PUERTOTERM |
| --- |
| *Si en una oración aparece un satélite como una relación de fondo entonces elimínela: IF (s) is satellite of a RESULTADO relation B Entonces si /A$^{prsa}$/>1 Eliminate S* |

Figura 60. 3 ra Regla PUERTOTERM

d) Justificación

| 4 ta Regla PUERTOTERM |
| --- |
| *Si en una oración aparece un satélite como una relación de fondo entonces elimínela: IF (s) is satellite of a JUSTIFICACION relation B Entonces si /A$^{prsa}$/>1 Eliminate S* |

Figura 61. 4 ta Regla PUERTOTERM

e) Circunstancia

| 5 ta Regla PUERTOTERM |
| --- |
| *Si en una oración aparece un satélite como una relación de fondo entonces elimínela: IF (s) is satellite of a CIRCUNSTANCIA relation B Entonces si /A$^{prsa}$/>1 Eliminate S* |

Figura 62. 5 ta Regla PUERTOTERM

### 5.9.2.- Criterios de supresión de núcleos discursivos

No se incluyen las relaciones discursivas Núcleo- Satélite cuando el texto describe asuntos relativos a Interpretación y Evidencia. Esto permitió desarrollar reglas para eliminar núcleos de Interpretación y Evidencia (Ver Anexo 12) (Figura 63-64).

**B) Se elimina el núcleo discursivo de las relaciones de: a) Interpretación y b) Evidencia.**

a) Interpretación

| 6 ta Regla PUERTOTERM |
| --- |
| *Si el Núcleo (N) de una relación de INTERPRETACION (I) entonces debe eliminarse N y deje el satélite de Interpretación. IF N is núcleos of INTERPRETACION them eliminate N and Keep S, I* |

Figura 63. 6 ta Regla PUERTOTERM

c) Evidencia

| 7ma Regla PUERTOTERM |
| --- |
| *Si el Núcleo (N) de una relación de EVIDENCIA (I) entonces debe eliminarse N y deje el satélite de Interpretación. IF N is nucleus of EVINDENCIA them eliminate N and Keep S, I* |

Figura 64. 7ma Regla PUERTOTERM

**C) No se separa el satélite de su núcleo en las relaciones de: a) Condición y b) Resumen** (Figura 65-66).

a) Condición

| 8va Regla PUERTOTERM |
| --- |
| *Si el Satélite (S) de una relación de CONDICION (C) entonces debe eliminarse N y dejar S en la relación de Condición. IF S is Satellite of CONDITION them eliminate S and Keep N, C* |

Figura 65. 8va Regla PUERTOTERM

c) Resumen

| 9na Regla PUERTOTERM |
| --- |
| *Si el Satélite (S) de una relación de RESUMEN (C) entonces debe eliminarse N y dejar el S en una relación de resumen. IF S is Satellite of CONDITION them eliminate S and Keep N, C* |

Figura 66. 9na Regla PUERTOTERM

D) **No suprimir ni separar los núcleos de las relaciones de: a) Contraste, b) Unión, c) Lista y d) Secuencia** (Figura 67-70).

a) Contrate

| 10 ma Regla PUERTOTERM |
| --- |
| *Si el núcleo está dentro de una relación de contraste no lo elimine IF N is nucleus of a CONTRASTE (C) ENTONCES NO LO SEPARE* |

Figura 67. 10 ma Regla PUERTOTERM

b) Unión

| 11 na Regla PUERTOTERM |
| --- |
| *Si el núcleo está dentro de una relación de UNION no lo elimine IF N is nucleus of a UNION (U) ENTONCES NO LO SEPARE* |

Figura 68. 11 na Regla PUERTOTERM

c) Lista

| 12 da Regla PUERTOTERM |
| --- |
| *Si el núcleo está dentro de una relación de LISTA no lo elimine IF N is nucleus of a LISTA (L) ENTONCES NO LO SEPARE* |

Figura 69. 12 da Regla PUERTOTERM

d) Secuencia

| 13 ra Regla PUERTOTERM |
| --- |
| *Si el núcleo está dentro de una relación de SECUENCIA no lo elimine IF N is nucleus of a SECUENCIA (S) ENTONCES NO LO SEPARE* |

Figura 70. 13 ra Regla PUERTOTERM

E) Reglas que eliminan elementos sintáctico-comunicativos

Las reglas que se presentan en este apartado están en dependencia de la sintaxis de dependencias (estructura profunda) estructura discursiva. La indiscutible supresión de elementos apenditivos (APPEND) y la posición preponderante de elementos coordinativos (COORD) en el resumen ha dado como resultado las siguientes reglas (Figura 71).

a) Se elimina un elemento sintáctico: Apenditivos.

| 14 ta Regla PUERTOTERM |
| --- |
| *Si la relación de dependencia aparece en la sintaxis de la oración como Apenditivo (APPEND) entonces elimine el apenditivo de la oración mencionada. IF in the sentence (s) have a appeditive element in the syntactic structure Them ELIMINATE APPEND* |

Figura 71. 14ta Regla PUERTOTERM

La presencia de elementos que poseen a la vez un determinado satélite de Elaboración que coincide con frases de relativo explicativas, no son incluidas en los resúmenes (Figura 72).

| 15 ta Regla PUERTOTERM |
| --- |
| *Si la oración es satélite de Elaboración (E) y posee una estructura sintáctica de dependencias como atributo de la oración entonces debe eliminarse la oración.* |

| |
|---|
| *IF the sentences (s) is satellite of an ELABORATION relation (E) and there is a syntactic dependency structure such that THEN ELIMINATE S* |

Figura 72. 15 ta Regla PUERTOTERM

b) No se suprimen aquellos satélites discursivos cuyos nexos con los siguientes elementos son inobjetables: Satélites de Elaboración correspondientes a elementos atributivos (en concreto, explicativos) (Figura 73).

| 16 ta Regla PUERTOTERM |
|---|
| *Si la oración es satélite de Elaboración (E) y posee una estructura sintáctica de dependencias como atributo de la oración entonces debe eliminarse la oración.*<br>*IF the sentences (s) is satellite of an ELABORATION relation (E) and there is a syntactic dependency structure such that THEN ELIMINATE S* |

Figura 73. 16 ta Regla PUERTOTERM

Son indiscutibles los nexos entre la organización discursiva y la ordenación comunicativa expresados en la posición que posee determinado satélite de discurso de Elaboración con referencia al elemento Tema del núcleo de la oración, tal es el caso que los autores deciden suprimir en sus resúmenes dichos elementos (Figura 74).

| 17 ma Regla PUERTOTERM |
|---|
| *Si la oración o sentencias (s) es un satélite de una relación de ELABORACION (e) y si la relación de elaboración que está en la oración es el tema del núcleo de Elaboración entonces eliminar la oración*<br>*IF S is satellite of ELABORATION (E)and (S) elaborates on the Theme of the nucleus (N) of E THEN ELIMINATE S (D´cunha, 2006)* |

Figura 75. 17 ma Regla PUERTOTERM

Se ha observado que si un satélite discursivo de Elaboración se refiere al Rema de su núcleo, los autores no lo usan en sus resúmenes (Figura 73).

c) No suprimir un satélite discursivo cuyos nexos con un elemento comunicativo sea evidente y probado, tal es el caso de los Satélites de Elaboración que se refieren al Rema de su núcleo.

| 18 va Regla PUERTOTERM |
| --- |
| *Si la oración es satélite de Elaboración y la oración de elaboración es el Rema del de Elaboración no elimine la oración. IF S is satellite of ELABORATION E AND S elaborates on the Rheme of the nucleus N of E THEN KEEP S(D´cunha, 2006)* |

Figura 76. 18 va Regla PUERTOTERM

### *5.10.- Criterios de desambiguación léxica*

Algo que no expone Iria D´Cunha (2006) en su trabajo son los criterios de desambiguación, o sea los elementos que se eliminan o suprimen en el texto cuando los autores van a desarrollar correcciones en el texto (Figura 77-79). Lo corroborado en los resúmenes que se construyeron en voz alta mediante la guía desarrollada a tales efectos (Ver Anexo 1y 26) es lo siguiente:

a) Conjugar todos los verbos de primera persona a voz pasiva.

| 19 na Regla PUERTOTERM |
| --- |
| *Todas las formas verbales que se encuentren en primera persona deben aparecer en voz pasiva. OII VERBAL FORMS IN PRESENT HAVE CONVERT IN PASIVE VOIX* |

Figura 77. 19 na Regla PUERTOTERM

b) Todas la referencias anafóricas del texto deben estar en la misma línea de cohesión (esto se aplica en la base de conocimientos) con pronombrs anafóricos.

| 20 ma Regla PUERTOTERM |
| --- |
| *Todas las referencias anafóricas deben estar en una misma voz lineal. OII ANAPHORIC REFERENCES HAVE REGULAR FORM* |

Figura 78. 20 ma Regla PUERTOTERM

c) Las puntuación en las oraciones son generalmente puntos y comas

| 21 ra Regla PUERTOTERM |
| --- |
| *Todos los signos de puntuación deben ser exactamente punto final y coma. OII ELEMENTS OF PUNTUATION ARE POIN OF COMA* |

Figura 79. 21 ra Regla PUERTOTERM

### *5.11.- Retos de Implementación de las Reglas Textuales*

La experiencia en la enunciación de reglas textuales en Cuba es muy joven, esta investigación es la primera que asume explícitamente estos saberes dentro del campo de la Ciencia de la Información o Documentación, esto hace que las miradas no recaigan tanto en los medios computacionales, si no, en los sistemas de representación de la información como un todo íntegro que facilite la asunción de modelos de análisis.

Las cuestiones que se exponen en este capítulo son los elementos básicos para la implementación del sistema, ya han sido declaradas las reglas para el trabajo con el texto, lo que supone nuevos retos en el desarrollo del software. Si bien en este capítulo se han establecido reglas sobre las cuestiones teóricas declaradas por D´Cunha (2006) y Mann y Thompson (1988) los aportes de las reglas que se llaman desde ahora PUERTOTERM logran desarrollar un sistema coherente sobre una base científica (desarrollada a partir del estudio de los textos), donde la impronta de las estudios de Mann y Thompson repercuten de manera teórica, para desarrollar nuevas reglas, acoger otras ya desarrolladas y aportar nuevos modelos de tratamiento del texto.

A las reglas desarrolladas por Iria D´Cunha en el 2006 se le agregaron algunas que aprovechan la textualidad y se contextualizan específicamente en los preceptos de PUERTOTERM; y otras, que particularizan los textos y los usuarios futuros del sistema que ha de diseñarse, cuestión esta última de un elevado nivel de complejidad, pero sin la cual no existe ningún sistema de información. El autor es consciente de que hasta aquí ha modelado al dominio lingüísticamente, falta la fase donde las demandas de información se expresen para que el sistema pueda entregar información oportuna, entonces, en forma de producto/servicio. Se trató de desarrollar reglas que no estuviesen en contraposición con los usuarios, pero también se desarrollaron otras formas de tratamiento de otros textos, que también están declarados en nuestro modelo y que permiten que el sistema sea más escalable y facilite el desarrollo de otros tipos de texto. En la misma línea de análisis se han declarado tres reglas de desambiguación léxica, consistentes exactamente en el tratamiento de las anáforas, los signos de puntuación y los verbos, se reconoce que estos

elementos no son suficientes para lograr la cohesión óptima, pero pueden ser debidamente adaptados al modelo de PUERTOTEX y a su estructura de software.

No se validan las reglas y las regularidades lingüísticas por ser procesos innecesarios para esta etapa. En el desarrollo del software se implementa un modelo de evaluación que incluye todos los procesos del sistema (Ver Capítulo 6).

Los retos esenciales de la programación son la búsqueda de un software y un lenguaje de programación, que ejecute este gran número de demandas textuales y cognitivas. El desarrollo de sistemas ontológicos reforzará el intenso trabajo del sistema al igual que la base de conocimientos. La futura implementación de estas reglas reviste altos niveles de dificultad, al igual que los procesos de evaluación de sistema, los cuales son complejos y muy costosos. En el plano del desarrollo de la aplicación la propuesta es elaborar solo aquellos elementos que sean necesarios para el desarrollo beta del proyecto y que permita evaluar el sistema y el modelo.


# Referencias Bibliográficas

### *5.12.- Referencias Bibliográficas*

ALVAREZ DE MON, I. 1999. *La Cohesión del texto .científico: un estudio contrastativo inglés-español.* Tesis Doctoral, Universidad Complutense de Madrid.

ATKINS, B. 1992. Tools for Computer-aided Corpus Lexicography: The Hector Project. *In:* KIEFER, F., KISS, G. & PAJZS, J. (ed.) *In Papers in Computational Lexicography. COMPLEX' 92.* Budapest: Linguistic Institute Hungarian Academy of Science.

BIBER, D., CONRAD, S. & SPENSER, R. 1998. *Corpus linguistics investigating language structure and use,* Cambridge, Cambridge University Press.

BOWKER, L. & PEARSON, J. 2002. *Working with specialized language,* Londres y New York, Routledge.

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

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

EDMUNDSON, H. 1969. New Methods in automatic extracting. *Journal of the Asociation of Machinery,* 16**,** 264-285.

ENDRES-NIGGEMEYER, B. 2005. SimSum: an empirically founded simulation of summarizing *Information Processing and Management,* 36**,** 659-682.

ENDRES-NIGGEMEYER, B., MAIRE, E. Y SIGEL, A. 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing & Management,* 31**,** 631-674.

FABER, P., MÁRQUEZ, C. & VEGA, M. 2005. Framing terminology: a process-oriented approach. *Meta,* 50**,** 189-213.

HERNÁNDEZ, A. 2006. indización y Resumen. La Habana: Universidad de la Habana.

HOEY, M. 1983. *On the Surface of Discoursa* London, George Alíen & Unwin.

LEIVA, A., SENSO, J., DOMÍNGUEZ, S. & HÍPOLA, P. 2009a. An Automat for the semantic processing of structured information. *In ISDA 9na International Conference of Desing of Software and Aplicactión.* Italia, Pissa: IEEE.

LEIVA, A., SENSO, J., DOMÍNGUEZ, S. & HÍPOLA, P. 2009b. An Automat for the semantic processing of structured information. *In:* In ISDA 9na International Conference of Desing of Software and Aplicactión, Italia, Pissa. IEEE.

LUNH, H. 1958. The Automatic creation of Literature abstracts. *Journal of Research of Development***,** 159 – 165.

MANN, W. & THOMPSON, S. 1990. Rhetorical structure theory: a theory of text organization.

MATHIS, B., RUSH, J. & YOUNG, C. 1973. Improvement of automatic abstracts by the use of structural analysis. *JASIS,* 24**,** 101-109.

MC ENERY, A. & WILSON, A. 1996. *Corpus linguistics,* Edimburgo, Edinburgh University.

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

MEL'CUK, I. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. *In:* AGEL, V., EICHINNGER, L., EROMS, H., HELLWIG, P., HERRINGER, H. J. & LOBIN, H. (eds.) *Dependency and Valency :an International Handbook of Contemporary Research.* Berlín - Nueva York: W. de Gruyter.

MOREIRO, J. 2006. *El resumen científico en el contexto de la teoría de la documentación. Texto y descripción sustancial* [Online]. Madrid. Available: Disponible en:http://www.ucm.es [Accessed 26.octubre 2006].

ONO, K., SUMITA, K. & MIIKE, S. 1994. Abstract generation based on rhetorical structure extraction. *Proceedings of the International Conference on Computational Linguistics.* Kyoto.

PANECA, F. 2009. *La Prensa remediana del siglo XIX: reflejo de la cultura popular tradicional de la localidad.* Tesis de Grado, Universidad Central "Marta Abreu" de las Villas.

PÉRES HERNÁNDEZ, M. C. 2002. *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento.* Málaga, Universidad de Málaga.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez

PINTO, M. 2004. Interdisciplinary Approaches to the Concept and Practice ofWritten Text Documentary Content Analysis. *Journal of Documentation,* 50**,** 405-418.

SALTON, G. 1996. On the application of syntactic methodologies in automatic text analysis. *Information Processing and Management,* 26**,** 73-92.

SÁNCHEZ, A., SARMENTO, T., CANTOS, P. & SIMÓN, J. 1995. *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones,* Madrid, SGEL.

SENSO, J. 2009a. *Representación del conocimiento en la Ingeniería de Puertos y Costas.* Proyecto Investigador, Universidad de Granada.

SENSO, J. & LEIVA, A. 2008. *Metamodelo para la extracción y desambiguación de textos científicos,* Santa Clara, Cuba, Universidad Central "Marta Abreu" de las Villas, Editorial Samuel Feijoó.

SENSO, J. A. 2009b. Proyecto investigador: Representación del Conocimiento. Granada, España: Universidad de Granada, Departamento de Documentación.


# Capítulo VI

## Puertotex: un software para la obtención de resúmenes en el dominio de Ingeniería de Puertos y Costas

# Capítulo 6: Puertotex: un software para la obtención de resúmenes en el dominio de Ingeniería de Puertos y Costas

**Introducción:**

Este capítulo describe los procesos de construcción de software y el servicio de información resultante de la aplicación. En este Capítulo de la investigación se muestran los elementos siguientes: Procesos de Marcación de Texto en XML, Diseño de la Base de Conocimientos, Construcción de la Ontología, Modelación de los Agentes e Ingeniería de Software y El diseño del servicio de Información aplicado al ámbito de la Universidad Central "Marta Abreu" de las Villas. Los elementos que se relacionan con la evaluación del software y su efectividad serán tratados en un capítulo específico en la tesis (Ver Capítulo 7).

## *6.1.- Implementación de las Reglas Textuales*

Según se declaró en el apartado anterior el primer proceso de implementación de las reglas es la detección de la unidades retóricas esenciales del artículo (I(O) MRC), considerándolos como textos autónomos para que la información más rica de cada uno de ellos sea utilizada para la extracción de texto. Para ello se desarrolla un script en Python[27] el cual funciona de la siguiente forma:

1. Identificación de la estructura retórica del artículo científico en Ingeniería de Puertos y Costas (Introducción, Objetivos, Metodología, Resultados y Discusión). Esto obliga al reconocimiento de las variantes diversas que ya han sido declaradas en el capítulo anterior.
2. A partir de la estructura retórica detectada: segmentar el texto en 4 apartados independientes insertando los textos en ficheros en documentos con extensiones txt. Por ejemplo Introducción.text.
3. Incluir las cuatro variantes de los subtítulos del artículo que se desea resumir en un fichero, con el nombre subtitulo.txt.

[27] El Scrip se denonima SATCOL y es realizado por Sandor Domínguez Velasco (Profesor de la Universidad de las Villas, Cuba)

4. Incluir el título principal del artículo en un fichero, con el nombre titulo.xml.
5. Guardar el título principal del texto en un fichero XML con nombre título.

### 6.1.1.- Propuesta de un conjunto de etiquetas XML en para el tratamiento de los niveles de estructura del texto

Para desarrollar los procesos de marcación de los elementos del texto se determina organizar las unidades textuales en segmentos. Dichos segmentos aúnan elementos de diversos órdenes como el discursivo, el sintáctico y el comunicativo, los cuales son referentes internos de las cargas semánticas y estructurales de cada oración. El desarrollo de estas etiquetas permite la obtención de resúmenes extractos de un texto único, por tanto, en este apartado de la investigación nos centraremos en el diseño de las etiquetas necesarias para estructurar los textos, que serán la entrada del sistema y que constituirán los elementos esenciales para su posterior evaluación automática y empírica.

Las concepciones en que se ha basado el autor para el diseño de las etiquetas obedece a tres criterios esenciales: la secuencia en que se etiquetan los niveles en la práctica, la simbología utilizada para declarar los elementos cohesivos del texto y la presencia en el texto de elementos geográficos, verbos, datos estadísticos, fórmulas y procesos. Las marcas o etiquetas están trazadas en XML y están integradas por un tag de inicio y un tag de salida Ej: <vb> Se estudian </vb>.

Lo primero que se marca es la oración que no estará prefijada por un punto, si no por la presencia de formas verbales y sus conectores. Luego de este paso se marca la estructura discursiva de cada sentencia de forma independiente, en otras palabras, se delimitan los núcleos y los satélites.

Cada relación discursiva aparece en el Capítulo 5 de la Investigación. Los nodos que se declaran y su forma se exponen en la tabla (Ver tabla 47). Puede observarse que en la columna de la derecha se muestran dos formas de marcación: etiquetas de relaciones Núcleo-Satélite y etiquetas de relaciones Multinucleares. Todas estas etiquetas tienen como inicio la letra <C>, lo que indica que se está en presencia de una relación de orden sintáctico y comunicativo. La marca que está después (<elab_n>) declara la tipología de la relación que se está marcando, en este caso elaboración, bien sea de núcleo o de satélite y dentro de él el número de relación. De esta forma la relación de elaboración establecida en la oración contiene la etiqueta que marca al núcleo y al satélite. Si existiese otra relación de elaboración tendría entonces el número 2 en las etiquetas de núcleo y en las de satélite.

Como puede verse todas las etiquetas que explican relaciones de tipo Núcleo-Satélite poseen una estructura similar, a excepción de los casos en que existe un satélite discursivo en relaciones de secuencia, lista o unión. En el caso de las presencia de satélites en la relación de Elaboración se enuncia en las marcas de discurso los elementos Tema y Rema de su núcleo. A igual que en los trabajos de (D'Cunha, 2006) esta construcción semántica se ejecuta a través del atributo <tem>, cuyo valor será el Tema o el Rema de su núcleo (marcados como "t" y "r", respectivamente, junto con el número con el que aparecen en su etiqueta), dependiendo de a cuál de los dos se refiera el satélite. Ej. <item="t">, <item="r">.

Es imposible encontrar satélites en relaciones multinucleares, esto obliga a solo declarar en la marcación los diversos núcleos, los que poseen igual nivel de significado para el resumen.

<C><lista><elab></elab></lista_n></C>

#### 6.1.1.1.- Relaciones de Comunicación

Las relaciones de comunicación se marcan con la letra C y están encargadas de establecer el sentido del texto a través de nexos entre las estructuras de núcleos y satélites (Ver Figura 46).

| Núcleo | Satélite | Relac. Discursiva |
|---|---|---|
| <C><elab_n></elab_n></C> | <C><elab_s></elab_s></C> | Elaboración |
| <C><inter_n></inter_n></C> | <C><inter_s></inter_s></C> | Interpretación |
| <C><evid_n></evid_n></C> | <C><Evid_s></Evid_s></C> | Evidencia |
| <C><fond_n></fond_n></C> | <C><fond_s></fond_s></C> | Fondo |
| <C><just_n></just_n></C> | <C><just_s></just_s></C> | Justificación |
| <C><rtad_n></rtad_n></C> | <C><rtad_s></rtad_s></C> | Resultados |
| <C><res_n></res_n></C> | <C><res_s></res_s></C> | Resumen |
| <C><condi_n></condi_n></C> | <C><condi_s></condi_s></C> | Condición |
| <C><refo_n></refo_n></C> | <C><refo_s></refo_s></C> | Reformulación |
| <C><circu_n></circu_n></C> | C><circu_s></circu_s></C> | Circunstancia |
| <C><lista_n><elab_n></elab_n></lista_n></C> | | Lista |
| <C><secu_n><elab></elab></secu_n></C> | | Secuencia |
| <C><uni_n><elab_n1></elab_n></uni_n></C> | | Unión |

Figura: 46 Relaciones de Comunicación

En segundo lugar, se etiqueta la estructura comunicativa, en términos de Tema y Rema. Las etiquetas empleadas se encuentran en la Tabla 47.

| Relación Discursiva | Tipo |
|---|---|
| <M><tem></tem></M> | Tema |
| <M><rem></rem></M> | Rema |

Tabla 47. Marcas de Tema y Rema

La primera etiqueta <M> se refiere a la forma en que se desarrollaría la relación que es de tipo Multifuncional, que se encuentra dentro de la estructura sintáctico-comunicativa. Después de la misma aparecen los nodos <tem>, <rem>. Estos tags evidencian que se declaran marcas referentes a Tema o Rema.

En este sistema de etiquetado se han declarado elementos de la superestructura del texto, lo que permitirá que el mismo sea reconocido por estándares como el Dublín Core, RDF y XML (Tabla 48).

| ***Super estructura*** | **Significado** |
|---|---|
| <dc:title></dc:title> | Título |
| <dc:creator></dc:creator> | Autor |
| <rdf:li><rdf:li> | Se encuentra dentro de la etiqueta autor e indica el nombre. |
| <dtc:afiliacion></dtc:afiliacion> | Entidad a la que pertenece El autor |
| <rdf:Description id=""> </rdf:Description id="001"> | Es el id del texto y por este será recuperado |
| <rdf:pu><a>3<no.>5<vol.>6</vol.></no.></a></rdf:pu> | Área de publicación y distribución, para explicitar los elementos ano, número y volumen. |

Tabla 48. Marcas en la Superestructura

A medida que el etiquetaje del texto avanza, la numeración de cada pareja de Tema y Rema irá aumentando. Es posible que en una oración no haya Rema, si no tema y entonces solo debe marcarse el Tema al que se le da el número secuencial que le corresponde. EJ: <M> <rem2> </tem2> </M>.

#### 6.1.1.2.- Relaciones Sintácticas

Finalmente se marca la estructura sintáctica de dependencias declarada con la letra E. La misma consta de los actuantes, relaciones apenditivas, atributivas, coordinativas y procesos. En la tabla se muestran la forma en que quedan las etiquetas y dentro de las mismas se declara la etiqueta text, para que el texto se insertado dentro de ella (Tabla 49).

| Relaciones Sintácticas | Formas de Relación |
|---|---|
| <E><act><text></text><act></E> | actuante I |
| <E><act2><text></text><act2></E> | actuante II |
| <E><act3><text></text><act3></E> | actuante III |
| <E><act4><text></text><act4></E> | actuante IV |
| <E><apend><text></text><apend></E> | Apenditiva |
| <E><atrib><text></text><atrib></E> | Atributivas |
| <E><fq><text></text><fq></E> | fórmulas químicas |
| <E><lug><text></text><lug></E> | Lugares |
| <E><stad><text></text><stad></E> | Estadística |
| <E><proc><text></text><proc></E> | Procesos |
| <S> <coord item="1"> <text> </text> </coord></S> | Coordinativas 1 |
| <S> <coord item="2"> <text> </text> </coord></S> | Coordinativas 2 |

Tabla. 49. Relaciones Sintácticas

Un elemento que se ha utilizado en esta tesis es el uso de articuladores, que sirve de elemento de entrada a las diversas oraciones, aunque no es un elemento comunicativo, el elemento se ha denominado <articulador> </articulador> ya que detecta los artículos en las oraciones. También se han incrementado elementos esenciales como lugares, estadísticas, procesos y fórmulas químicas, que son elementos que si bien no indican relaciones, declaran nexos entre diversos elementos dentro de los grupos oracionales.

### 6.1.1.3.- Marcado de las Relaciones de Comunicación

Como se ha dicho en los acápites anteriores los textos deben marcarse a nivel oracional y discursivo. No existe ninguna herramienta para el marcado de relaciones discursivas en el español que facilite la marcación de núcleos, satélites, temas y remas en los textos. Con el objetivo de desarrollar de forma semiautomática este marcado se ha decidido desarrollar una interfaz de anotación utilizando como herramienta a XMLMarker, herramienta capaz de administrar bases de datos en XML, además con MySQL (Xampp o Lampp) y un script capaz de crear, formular y eliminar oraciones y registros es posible manipular archivos, etiquetar de forma semiautomática los textos, además de

ejecutar sentencias en XML y en Python, este último lenguaje que pude ser perfectamente compatible con Python.

Las acciones que deben acometerse para etiquetar un texto en la herramienta den ser las siguientes:

- Localizar la dirección http://xmlmarker.server/ (debe tenerse instalado XML Marker).
- Poner el usuario y la contraseña correspondiente sobre la carpeta de los textos.
- Seleccionar el texto que se desea resumir, en esta investigación el texto debe subir de forma manual debido a que las formas de construir las relaciones ha hecho muy engorrosa la subida automática de los textos.
- Declarar los elementos de la superestructura del texto: autor, título y filiación de acuerdo a lo que indica RDF.
- Etiquetar las oraciones del texto. Este procedimiento exige las siguientes acciones:

  a) Seleccionar la opción texto completo para ver todas las oraciones que se refieren a un texto específico.
  b) Leer oración inicial
  c) Leer oración subsiguiente
  d) Determinar la relación oracional
  e) Si está en presencia de una relación multinuclear declare las diversas relaciones de acuerdo con la simbología.
  f) Insertar los datos necesarios para el marcado de la oración
     - Número de la oración que ha de marcarse ‹ora_1› con su tag de cierre ‹/ora_1›
     - Insertar el Texto oracional.
     - Insertar las relaciones discursivas.
  g) Leer la oración subsiguiente
  h) Determinar la existencia de relaciones con relaciones antecedentes e identificar la relación que existe entre ellas.

i) Realizar el paso e hasta haber etiquetado todas las oraciones del texto.

j) Guardar el texto en el servidor asignándole una URL para que luego la ontología direccione esta posición de modo que el agente de resumen y el de búsqueda puedan leer el texto y resumirlo.

Concluido el proceso de marcado se genera un árbol de jerarquías, el que servirá de entrada para la aplicación de las reglas PUERTOTERM (Ver figura 80).

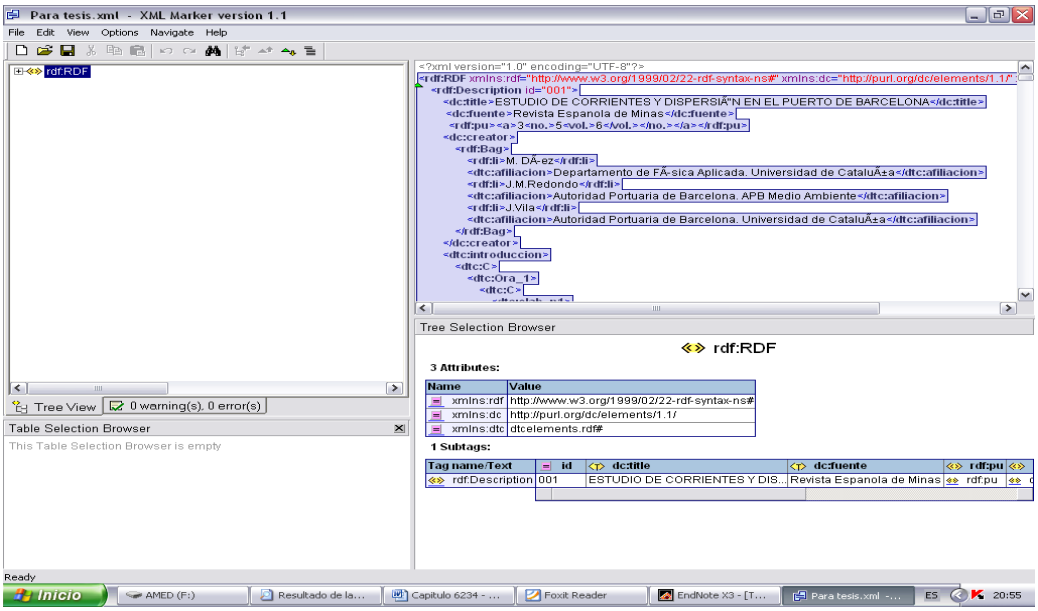



Figura. 80. Marcación

## *6.2.- Formalización de la Estructura*

La formalización de la estructura de los valores de la Ontología es un proceso que determina la estructura que tendrá la salida de la ontología, así como los procesos que servirán para la consulta. Los elementos que se utilizan para determinar la estructura en este trabajo están sujetos al análisis del RDF, sus aplicaciones y sus posibilidades de sintaxis.

### 6.2.1.- RDF

El RDF (Resource Description Framework) (Lassila and Webick, 1998) es una estructura de datos generados como un estándar por el W3C. RDF define la sintaxis y la semántica de un conjunto de datos y documentos para el entorno Web. Con RDF se logra la interoperabilidad en los sistemas y la escalabilidad de las herramientas que se basen en él. Este estándar permite formalizar tres tipos de representaciones de los datos:

- Tripletes
- Grafo
- XML

Debido a la existencia de nexos en las representaciones aquí se ha determinado trabajar con RDF, ya que las posiciones que asumen cada una de las formas de trabajo con RDF no limita su semántica, ni su expresividad, al contrario, al trabajar con los tres modos anteriormente definidos se eleva la capacidad del sistema para ser escalable . El RDF concretamente se estructura como se explica a continuación. Este estándar posee dos conjuntos básicos de elementos: Resource y Literals, dentro del conjunto Resource existe un subconjunto de datos denominados Properties. Los tripletes se conforman a partir del conjunto de Statements en el que los elementos del triplete se definen como (pred., sub., obj.), en el cual pred es un elemento de Properties, sub. Es un Resource y obj., puede ser o un Resource o un Literal.

Con RDF, desde el punto de vista sintáctico se logra que los sucesos sean tripletes anexados a Statements. Esto hace que el Statement base conserve su status tal y como enuncia Fernández Bréis (Fernández Breis, 2003) sigue siendo un suceso a pesar de ser reificado, puesto que el triplete que representa al Statement original permanece en Statements. Con esto se deja claro que RDF se guía por tripletas con una estructura (sujeto, predicado, objeto).

El sujeto es un elemento que se idéntica con una URI, y se relaciona con el objeto mediante un predicado binario que puede ser un elemento URI o un literal. RDF basa su sintaxis en tres posibilidades específicas:

- XML[28]
- N3[29](notación de tripletas)
- Turtle[30]

Todas tienen un valor sintáctico, en dependencia de lo que se desee obtener en la programación. Desde la óptica de este proyecto se ha determinado usar N3. XML posee un sofisticado aparato de tratamiento de la información, sin embargo, sus potencialidades para la consulta obligaban a eliminar algunos elementos en el proceso de búsqueda de la información. N3 por su parte es más sintáctico y expresivo si se quieren realizar consultas múltiples. Turtle, aunque posee una elevada reputación, por su alto nivel de implementación es una definición sintáctica que reducía la semántica de los registros, por tanto nos obligaba a hacer cambios en las concepciones del software (Figura 81).

**Algunos Elementos del RDF**

```
owl:Class
rdf:about="http://www.semanticweb.org/ontologies/2010/4/Ontology1274835264580.owl
#AGUA">
   <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/2010/4/Ontology1274835264580.
owl#Agentes_Naturales"/>
   <dc:subject>BOSQUES</dc:subject>
   <dc:subject>RIOS</dc:subject>
   <dc:title>Laguna Mata Redonda
</dc:title>
   <dc:subject>PROTECCION DEL CLIMA</dc:subject>
   <dc:title>Agua, recurso estrat&#233;gico garantizado en el Estado de M&#233;xico
</dc:title>
   <rdfs:seeAlso>LAGUNA</rdfs:seeAlso>
   <dc:title>Aguas salvajes y de arroyada</dc:title>
   <tipo_de_articulo>Semi-Especializados</tipo_de_articulo>
   <dc:creator xml:lang="pt">
```

Figura: 81. Elementos de RDF

### 6.2.1.2.- Vocabulario RDF

RDF es una ampliación de OWL (Web Ontology Language) protocolo instituido por W3C[31] para el desarrollo de ontologías en la Web (Corrales del Castillo,

[28] http://www.w3.org/TR/rdf-syntax-grammar/
[29] http://www.w3.org/DesignIssues/Notation3
[30] http://www.dajobe.org/2004/01/turtle/
[31] http://www.w3.org/

2008) . OWL es un lenguaje de alto nivel semántico con una amplia capacidad para la descripción de conceptos y relaciones conceptuales, además facilita el tratamiento semántico de cada entidad o dato. Es OWL, la versión superior de los antiguos lenguajes de tratamientos de ontologías (DAML[32], DAML[33] + OIL).

Son muchas las aplicaciones de los vocabularios RDF, entre ellas se encuentra EARL, DOAP, Dublín Core, RSS, FOAF, etc. En esta investigación solo se utilizan Dublín Core y FOAF. A continuación se describe cada uno de ellos dejando al final los utilizados en la tesis.

**EARL (Evaluation and Report Language):** es un léxico en RDF que facilita el registro, suministro y procesamiento de datos sobre evaluaciones automáticas. Su uso ha estado asociado a los protocolos de acceso sobre todo en TAW[34], con el fin de exponer reportes de los procesos evaluativos.

**DOAP (Description-of-a-Project):** es un proyecto que posee similaridades con **FOAF**, pero su cometido es la descripción de todo proyecto de investigación.

**RSS:** está amparado por Netscape, y se ha convertido en un formato de sindicación de contenidos muy usado en páginas Web. Su gran defecto es la poca escalabilidad del sistema ya que sus diversas versiones son incompatibles entre sí (Figura 82).

Los vocabularios que se usarán en esta aplicación para usar el RDF se han seleccionado de acuerdo a las necesidades del proyecto.

**Dublin Core:** Dublin Core[35], también conocido por sus siglas DC, es un vocabulario RDF para la descripción de múltiples propiedades de todo tipo de recursos online. Los elementos que posee Dublín Core son los siguientes:

[32] http://www.daml.org/
[33] http://www.daml.org/2001/03/daml+oil-index
[34] http://www.tawdis.net/
[35] http://dublincore.org/

**Contenido:**

- **Título:** el nombre dado a un documento o recurso Web. Es un dato que le otorga generalmente el autor con el fin de identificarlo. *Etiqueta: DC.Title*

- **Claves:** Son los temas que tratan los recursos intrínsecamente. Para lograr la formalización léxica de los contenidos se utilizarán los conceptos debidamente declarados en la Ontología. *Etiqueta: DC.Subject*

- **Descripción:** Es una descripción resumida del recurso. *Etiqueta: DC.Description*

- **Fuente:** Es un conjunto de caracteres determinados para identificar los recursos de información que han de describirse. Etiqueta: *DC.Source*

- **Lengua:** lengua en que se expresa el contenido del recurso. *Etiqueta: DC.Language*

- **Relación:** Es un envío a un segundo recurso en el sistema, en el caso nuestro sirve para cargar el documento en XML desde la ontología. *Etiqueta: DC.Relation*

- **Cobertura:** SE refiere a la cobertura, espacial y temporal del contenido de un recurso. Etiqueta: *DC.Coverage*

**Propiedad Intelectual:**

- **Autor o Creador:** La persona o organización que se adjudica el contenido intelectual del recurso. Etiqueta: *DC.Creator*

- **Editor:** Persono o entidad encargada de que un recurso esté disponible. Etiqueta: *DC.Publisher*

- **Otros Colaboradores:** Persona o entidad que esté relacionada con el contenido intelectual del recurso entre ellos está: editor, ilustrador y traductor. *Etiqueta: DC.Contributor*

- **Derechos:** Sirve para declarar los derechos de autor, generalmente constituye una nota. Etiqueta: *DC.Rights*

**Instanciación:**

- Fecha: fecha de creación del recurso. Etiqueta: *DC.Date*

- **Tipo del Recurso:** Tipología de cada recurso. Por ejemplo, página personal, romance, poema, diccionario, etc. Etiqueta: *DC.Type*

- **Formato:** Sirve para declarar el software y el hardware necesarios para acceder a un determinado recurso. *Etiqueta: DC.Format*

- **Identificador del Recurso:** Dirección física del recurso generalmente la URL o el ISBN Etiqueta: *DC.Identifier*

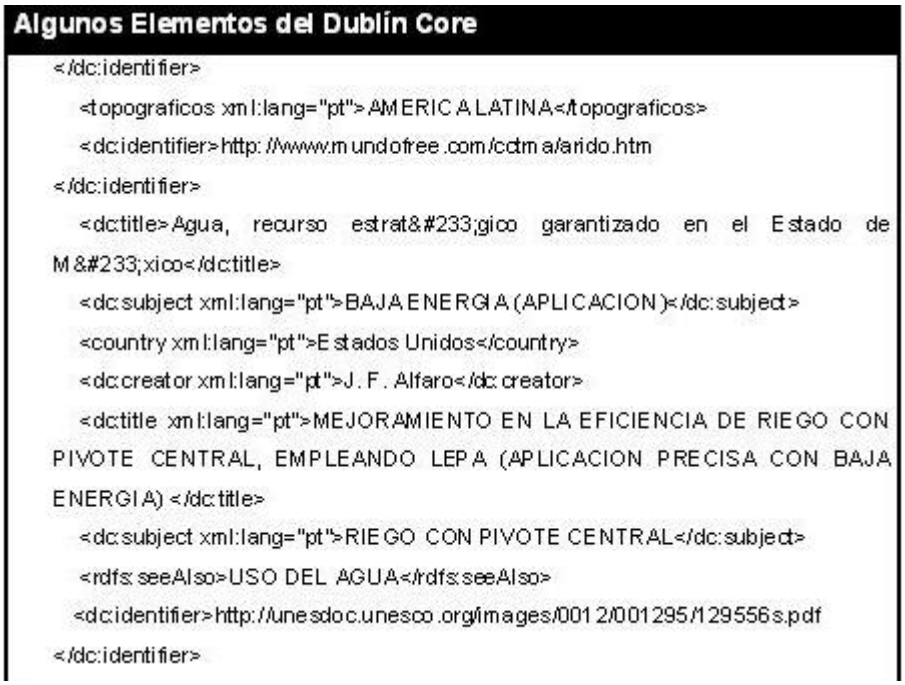



Figura: 82. Elementos de Dublín Core

**FOAF** (Friend-of-a-Friend) es un léxico de un vocabulario RDF para delimitar semánticamente información personal (83-84). Muy extendido en los sitios de redes sociales con Faceboock, etc. Foaf posee las siguientes clases:

- **Elementos de FOAF Básicos** (Se detallan todos los elementos para nombrar e identificar una persona), los elementos que pueden verse aquí son: *agent, person, name, nick, title, homepage, mbox, img, surname.*
- **Cuentas en Línea y Mensajería** (identifica las diferentes cuentas on line que posee una persona, así como los servicios de mensajería instantánea a los que accede).
- **Grupos y Proyectos:** Identifica a las personas que están en un proyecto y a sus financistas (fundedBy, Group).
- **Información Personal** (facilita el acceso a páginas personales e institucionales donde aparece información sobre los que se integran a un proyecto).
- **Documentos e Imágenes** (Facilita el acceso a imágenes y a documentos de personas que están en una red social).

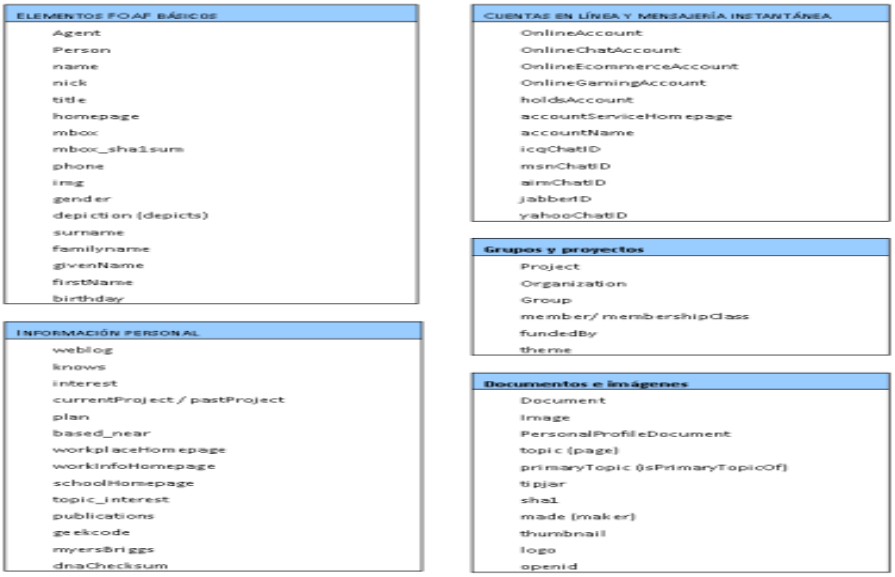



Figura: 83. Estructura del FOAF

**Algunos Elementos de FOAF**

```
!--
http://www.semanticweb.org/ontologies/2010/4/Ontology1274835264580.owl#MACHADO --

 <owl:NamedIndividual
rdf:about="http://www.semanticweb.org/ontologies/2010/4/Ontology1274835264580.owl#MACHADO">

   <rdf:type rdf:resource="&foaf;firstname"/>

 </owl:NamedIndividual>
```

Figura: 84. FOAF en RDF

Las limitaciones que posee RDF son esencialmente semánticas, pues no permite construir relaciones que no sean de tipo IS-A. Interacciones conceptuales que presenten disyunción, negación, etc., no pueden ser desarrolladas con este lenguaje a no ser que se hagan formalizaciones y restricciones individuales en el editor de la ontología.

#### *6.2.1.2.1.- RDF: sintaxis*

Los documentos RDF están compuestos de una serie de componentes que hacen que estos instrumentos posean peculiaridades que a continuación se describen:

- Recursos, subconjunto de atributos a los que se le denomina “propiedades” relacionado a un conjunto de literals.
- La sintaxis se construye a partir de estructuras de tripletas (sujeto-objeto y predicado.
- Todos los recursos constituyen instancias de la clase *rdfs: Resource* y su descripción está sujeta a las capacidades de *rdfs: Description,* la cual posee todos las propiedades y los valores correspondientes, además los elementos descritos poseen una referencia URI, expresada mediante el atributo *rdf: abaut,* el cual puede combinarse con rdf: ID.
- Los nodos anónimos se pueden referenciar mediante *rdf: Description* conjuntamente con *rdf: nodeID,* el que permite referenciar cualquier objeto en el documento RDF.

- Las clases y las subclases pueden comportarse al mismo tiempo como una u otra cosas, es decir una clase pude ser clase y subclase en un momento dado. Las propiedades que lo describen son *rdfs: Class y rdfs: Subclass.*
- La propiedad *rdf:Type* define el tipo de recurso, asignándole una URI
- La reificación de los elementos se realiza mediante las siguientes propiedades: *rdf: subjet*, *rdf: predicate* y *rdf: objet.*
- El lenguaje RDF utiliza contenedores que son capaces de unir determinados recursos con valores de una propiedad de pertenencia dada. Los contenedores que más se observan en RDF son *rdf: Bag* (para el que el orden de los recursos no es relevante), *rdf: Seq, (*el cual si se detiene a analizar el orden de los recursos) y rdf: Alt (permite seleccionar un recurso y excluirlo).
- Las propiedades del RDF se construyen mediante *rdf: Property,* utilizada para declarar aspectos que caracterizan a varias clases. También es posible dar propiedades más específicas de las clases a través de *rdf: Subproperty.* Las propiedades individuales se construyen mediante rdf: *DataType.*
- RDF: Posee amplias propiedades de procesamiento gracias al atributo *rdf: parseType,* capaz de tomar diversos valores entre los que se encuentra: *Literal, Resource y Collection*.
- Con RDF es posible determinar rango y domino, es decir se puede declarar relaciones específicas en las clases mediante *rdf: range* y *rdf: domain*.
- Un aspecto importante del RDF son las propiedades de definición facilitadoras de información para el usuario, ellas son: rdf: label (permite decir en qué idioma está el registro o el recurso), *rdf: comment* (facilita añadir información adicional para el usuario) y *rdf: seeAlso* (permite añadir información nueva al recurso que se describe). Es importante destacar que *seeAlso posee una subpropiedad* denominada *rdf: isDefinedBy,* la cual permite definir la semántica de un recurso RDF.

### 6.1.3.- Diseño de la DTD

Con el objeto de facilitar el marcado de los grupos oracionales para la construcción de los resúmenes por extracción se desarrolló una DTD (Document Type Defintion) (ver Anexo 36). La DTD utilizada para esta investigación es una norma que define claramente los elementos sintácticos que se necesitan para etiquetar los textos de nuestro corpus. El principal cometido de esta DTD es la descripción de la estructura de los datos, para poder homogenizar el tratamiento de los textos en el momento del marcado. La DTD quedó de la forma siguiente:

a) elementos (delimitan las etiquetas utilizadas y su contenido).
b) estructura (define el orden en que debe estructurarse el texto y en el que van los elementos de marcado).
c) anidamiento (permite determinar aquellas etiquetas que salen unas de otras).

Veamos, a modo de ejemplo, un fragmento de nuestra DTD, que describe algunos de los elementos de nuestra lista de elementos posibles:

!ELEMENT E (act1?, act2?, act3?, act4?, apedn?, atrib?,fq?, lug?, stad?, proc? , coord, E*)>

<!ELEMENT act1 (text)>

<!ELEMENT act2 (text)>

<!ELEMENT act3 (text)>

<!ELEMENT act4 (text)>

<!ELEMENT apedn (text)>

<!ELEMENT atrib (text)>

<!ELEMENT fq (text)>

<!ELEMENT stad (text)>

<!ELEMENT text (#PCDATA)>

A la DTD también se le añaden otros elementos que proceden de los estándares Dublín Core y RDF de los que se exponen sus características (Anexo 36). Un ejemplo de la DTD es mostrada en el siguiente ejemplo, ver figura 85).

**Algunos Elementos de la DTD**

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dtc="dtcelements.rdf#">

<rdf:Description id="001">

<dc:title>ESTUDIO DE CORRIENTES Y DISPERSIÃ"N EN EL PUERTO DE
BARCELONA</dc:title>

<dc:fuente>Revista Espanola de Minas</dc:fuente>

<rdf:pu><a>3<no.>5<vol.>6</vol.></no.></a></rdf:pu>

<dc:creator>

<rdf:Bag>

<rdf:li>M. DÃ-ez</rdf:li>

<dtc:afiliacion>Departamento de FÃ-sica Aplicada.
Universidad de CataluÃ±a</dtc:afiliacion>

<rdf:li>J.M.Redondo</rdf:li>

<dtc:afiliacion>Autoridad Portuaria de Barcelona. APB
Medio Ambiente</dtc:afiliacion>

<rdf:li>J.Vila</rdf:li>
```

Figura: 85. DTD

### *6.2.- Bases de Conocimiento en Puertotex*

Para desarrollar la implementación de las Reglas Textuales en el software Puertotex se han implementado dos bases de conocimiento. Estos sistemas de

inteligencia han sido definidos en la literatura de formas diversas y a veces son tomados por muchos autores como OECD (OECD, 2000) como sistemas de archivo de información en ordenadores. Esta definición de OECD (OECD, 2000) asemeja a las bases de conocimiento con las base de datos. Las últimas constituyen una parcela de la Ciencia de la Computación que ha recibido muchas miradas desde la Ciencia de la Información, la Ingeniería de Software, la Lingüística, la Recuperación de la Información y la Economía de Empresas. Las bases de datos pueden ser estructuradas y no estructuradas. Las bases de datos estructuradas recogen información que posee alguna estructura, o sea almacenan información estructurada. La organización de estos datos se logra mediante la sintaxis establecida a partir de los lenguajes de programación. Las bases de datos no estructuradas son a su vez sistemas que mueven datos cuya estructura varía. Las bases de datos no son sistemas de inferencia de conocimiento, sus datos solo se usan para recuperar información y no son capaces de inferir por sí solas conocimiento ya que no incluyen inteligencia ni razonamiento en su estructura y concepción.

### 6.2.1.- Base de Conocimiento posicional

Gracias a los desarrollos de la Inteligencia Artificial (IA) han aparecido Sistemas Basados en Conocimiento (SBC), cuyos aportes prácticos no han sido suficientemente tangibles, considerándose primitivos desde el punto de vista de concepción. Sin embargo las aplicaciones de la (IA) en diversas áreas de la vida cotidiana y la gestión de información evidencian la utilidad de estas técnicas de tratamiento de conocimiento. Los pasos para desarrollar nuestro Sistema Basado en Conocimiento se basa en los siguientes pasos:

- **Delimitación del Dominio:** Consiste en determinar los límites del domino de Ingeniería de Puertos y Costas.
- **Acopio del Conocimiento:** Adquisición del Conocimiento estructurado mediante técnicas de selección de términos. Estructuración de los contenidos mediante mapas conceptuales
- **Integración del Sistema**: Construcción en XML de una base de conocimiento con las heurísticas

- **Verificación y Validación del Conocimiento:** Conocimiento validado por el dominio**.**
- **Actualización de la Base de Conocimiento:** Actualización de la base de conocimiento para el uso de los contenidos.

Todos estos paso son realizados por un equipo de trabajo altamente calificado que se encarga de la estructuración del SBC, el cual integra toda la experiencia de los especialistas del campo de la Ingeniería de Puertos y Costas. Cada una de estas actividades es desarrollada principalmente por personal altamente calificado y, naturalmente, experto en la instalación de SBC. El Acopio de Conocimiento es la parte de más dificultad en el proceso de adquisición del conocimiento, pues el acto de extraer experiencia del conocimiento de un experto humano obliga al establecimiento de técnicas efectivas para determinar la estructura del mismo. Desde el punto de vista metodológico es innegable la existencia de propuestas metodológicas para abordar el fenómeno, las cuales son cotosas, ya que exigen el empleo a fondo de expertos y programadores.

Desde la óptica de Puertotex los procesos de desarrollo de la base de conocimientos se describen a partir de técnicas de selección de términos y de corpus, los procesos de validación de estrategias cognitivas, y la implantación del sistema, etapas que serán abordadas a continuación con detalle.

#### 6.2.1.1.- Base de Conocimiento (Construcción de Relaciones)

La necesidad de desarrollar una Base de Conocimiento dentro la herramienta obedece a la necesidad de contar con un instrumento para describir el conocimiento generado en este dominio y sus interacciones. Las bases de conocimiento que se definen en este proyecto tienen un enfoque orientado al dominio, es decir, están basadas en un dominio muy estructurado donde el razonamiento se basa en una heurística (estrategias para resumir y para consultar y denotar información), por tanto es un razonamiento basado en casos seleccionados del comportamiento de los usuarios. Esto se tornaba en un problema de implementación ya que la herramienta que se utilizó para desarrollar la ontología (protégé) no podía reconocer relaciones particulares ni

inferir conocimiento a partir de propiedades de datos y objetos ordenadas alfabéticamente en un fichero OWL. Esta situación obligó a desarrollar una base de datos en MySQL Server por la aplicación PhpMyadmin, en la cual pudiera escribirse un código en RDF en el momento de la consulta del usuario por medio de un agente. Para recopilar el conocimiento acopiado en un fichero Excel se construyó un pequeño comando en PERL, que permitió exportar la base de datos a MySQL. De MySQL se exportó la base de datos a XML y se creó una función for para escribir en un fichero texto cada una de las relaciones conceptuales, utilizando como lenguaje de programación Python, un lenguaje de alto nivel orientado a objetos. La base de conocimientos se llama Puertotex.db y posee tres tablas cada una con dos campos. A continuación mostraremos las tablas con sus respectivos campos (Ver tabla 50).

| Tabla | Campo Llave | Campo Sec. |
|---|---|---|
| Dominio | Dominio_id | Dominio |
| Relación | Relacion_id | Relación |
| Rango | Rango_id | Rango |

Tabla 50. Campos

Debido a que las relaciones, los dominios y los rangos pueden repetirse, la base de conocimientos fue normalizada a fin de que no hubiese datos repetidos y que no se produjeran errores conceptuales en el sistema de bases de conocimiento. Esta base de datos comprueba las restricciones creadas en la ontología y las escribe en un fichero RDF de forma automática para facilitar la construcción de relaciones en el momento en que se necesite realizar una búsqueda o una visualización.

#### 6.2.1.2.2.- Selección de Términos

La selección de términos y de los corpus es un proceso esencial en la ordenación de conocimiento léxico, para ello se apela al uso de la herramienta FoxCorp (Domínguez, 2011b), diseñada para este proyecto por no existir herramientas al uso en el mercado para marcar textos en inglés y español, integrando sus resultados con XMLMarker. Para evaluar la selección la calidad

de los términos se utilizaron diferentes medidas de evaluación, las cuales se integraron a una metodología particular capaz de evaluar los textos y los términos que iban a integrar la futura base de conocimiento (Capítulo 7).

## *6.3.- Ontosatcol: una ontología para el dominio de la ingeniería de Puertos y Costas*

### 6.3.1.-Estructura

Ontoclas es una ontología diseñada para la desambiguación, extracción y búsqueda de información dentro de la herramienta Puertotex. Este sistema ontológico cuenta con 3005 conceptos, 3006 subconceptos y 28900 anotaciones. Posee también 12 propiedades individuales para cada objeto y 8 propiedades especiales para los datos. La herramienta tiene además 890 instancias. Todos los términos y las instancias se han obtenido a partir de las bases de datos del Proyecto Puerto Term de la Universidad de Granada. Es importante destacar que los términos que aquí aparecen son obtenidos mediante los procesos de extracción de términos a través de la herramienta World Smith Tolls y seleccionados a partir de su frecuencia de aparición en cada texto y validados por los expertos del proyecto Puertoterm. Desde el punto de vista idiomático los conceptos que se describen aquí pertenecen al inglés y al español (Ver figura 85).

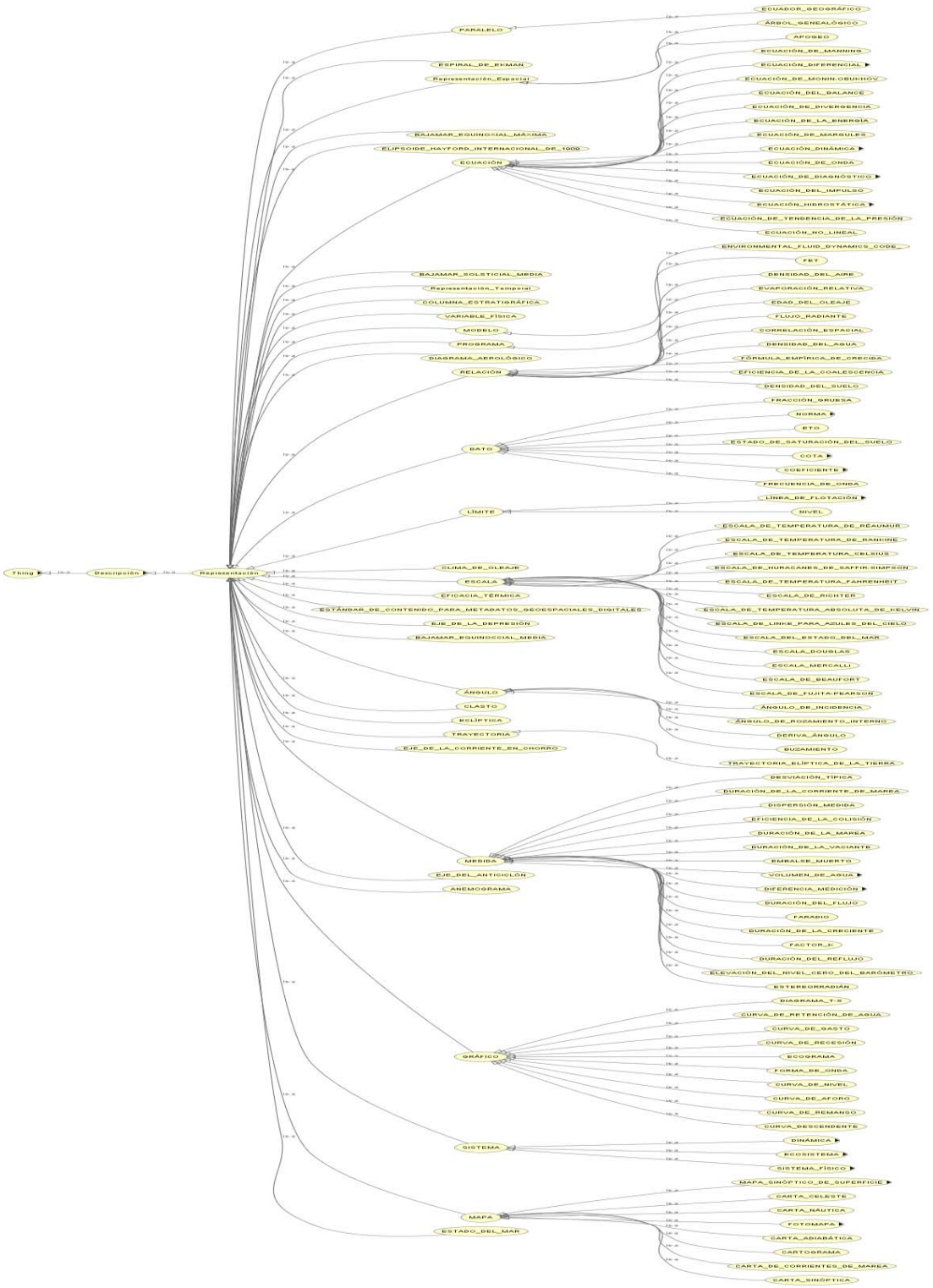



Figura. 85 Estructura arbórea de la ontología

### 6.3.1.1.- Clases y Subclases de la Ontología

La ontología está integrada por 4 clases generales declaradas a partir de los elementos que integran los procesos de actuación en el terreno de la Ingeniería de Puertos y Costas (ejes semánticos). Las clases en Ontosatcol están

dispuestas en forma jerárquica, así es posible desarrollar una estructura donde los Agentes, tengan modos de actuación descritos (Descripción) que intervengan en los procesos de actuación de dominio ya sean de orden artificial, natural o una posible causa y efecto de un fenómeno (Ver figura 86).

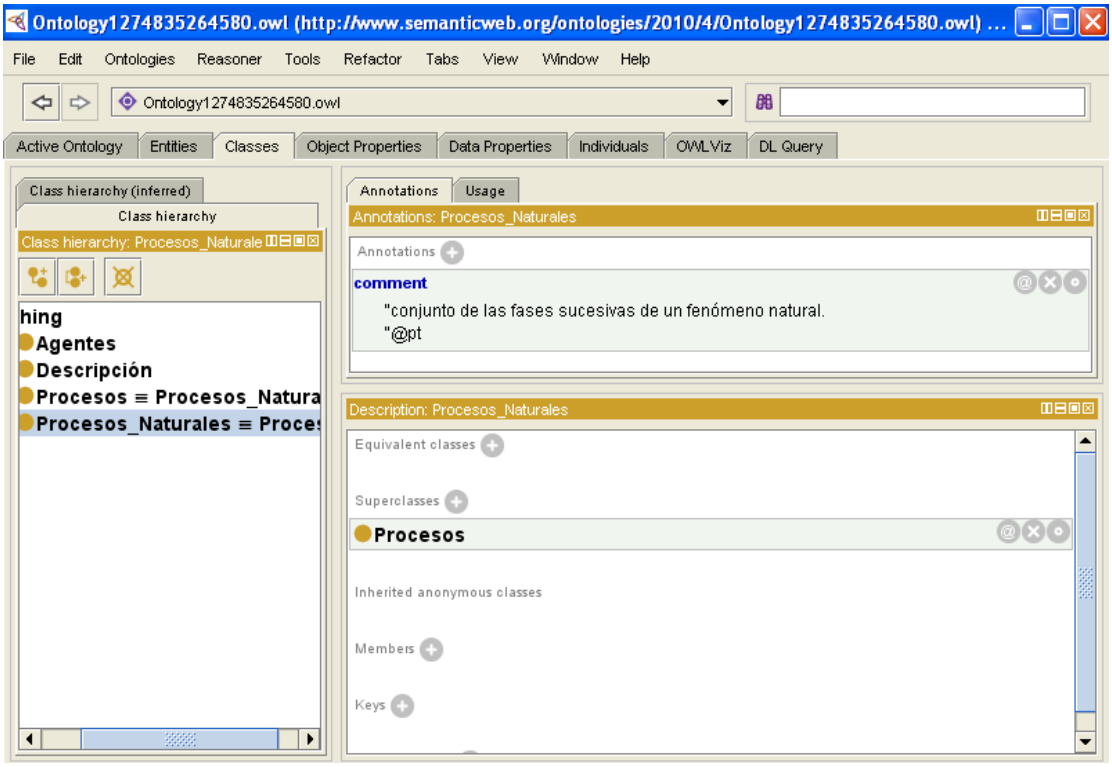



Figura: 86. Clases

Las subclases están concebidas en forma de jerarquías específicas (ver figura), es por ello que se detallan los agentes (Naturales y Artificiales), en la descripción se muestran las subclases: Atributos y Medidas (expresan la forma en que se mide algún elemento o fenómeno), Representación (denota los instrumentos de representación de los fenómenos) y por último Disciplinas de Estudio (denotan las especialidades que estudian los fenómenos). También se declaran las subclases de la clase Procesos que se subdivide en Procesos Naturales y Procesos Artificiales (Ver Figura 87).

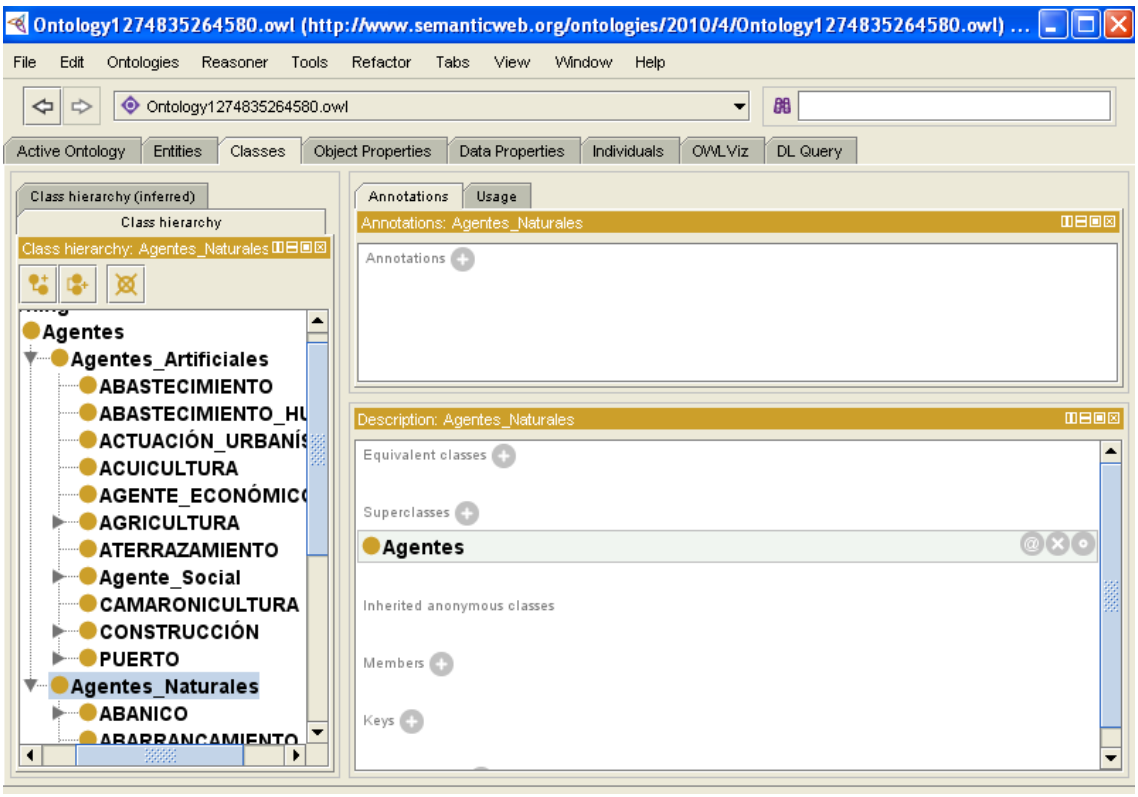



Figura: 87 Subclases

### 6.3.1.2.- Propiedades de los Objetos

Las propiedades de los objetos son 12 y han sido declarados por los expertos del proyecto Puertoterm. Cada propiedad tiene una semántica particular que se detalla a continuación:

- Afecta a: Todo evento que afecte a otro.
- Atributo de: Se incluyen aquí aquellos elementos que sean propiedades de un fenómeno y organismo.
- Causa: Referido a lo que se origina a partir de una situación o fenómeno.
- Contrario de: Denota lo opuesto de un fenómeno o suceso.
- Delimitado por: Concepto que delimita otro.
- Estudia: Se usa para disciplinas de estudio
- Parte de: Se utiliza cuando un concepto es parte de otro.

- Representa: Es formalizado para aclarar cuando hay un fenómeno o concepto o atributo que puede ser representado mediante un instrumento.
- Resultado de**:** Especifica que un elemento es resultado de otro.
- Se hace con: Sirve para declarar con que se hace determinada medición o actividad.
- Tipo de: Es para declarar un variante o especificidad de un concepto.

Estos objetos tienen tres únicas propiedades (Transitivas, Reflexivas y Condicionales) que son declaradas en las siguientes definiciones:

- Transitiva: $\forall a, b, c \in A,\ (a, b) \in R \wedge (b, c) \in R \ \Rightarrow\ (a, c) \in R$
- Reflexiva: $\forall x \in A,\ (x, x) \in R$
- Funcional: Dada una función $f : X \rightarrow Y$ es inyectiva cuando se cumple alguna de las dos afirmaciones equivalentes: Si $x_1$,$x_2$ son elementos de $X$ tales que $f(x_1) = f(x_2)$, necesariamente se cumple $x_1 = x_2$. Si $x_1$,$x_2$ son elementos diferentes de $X$, necesariamente se cumple $f(x_1) \neq f(x_2)$

### 6.3.1.3.- Propiedades de los Datos

Las propiedades de los datos en esta ontología son de tipo reflexiva es decir que un concepto puede ser visto como rango y como dominio a la misma vez (Ver Figura 87). Los elementos léxicos que declaran las relaciones entre los datos son los siguientes:

- **Compuesto de (material):** Indica la composición de algún material.
- **Fase de:** Define que conceptos constituyen un paso o fase de un proceso.
- **Onomásticos:** Utilizado como dato especial cuando un texto se refiere a una persona específica, se utiliza como una subdivisión de materia dentro de los datos del Dublín Core.
- **Tiene Función:** Identifica la función de determinada actividad o cosa en un procesos determinado.

- **Topográfico:** Es un dato utilizado en el registro del Dublín Core para declarar el lugar de algo.
- **Ubicado en:** Se refiere a la existencia u ocurrencia de un fenómeno o proceso en determinado lugar.

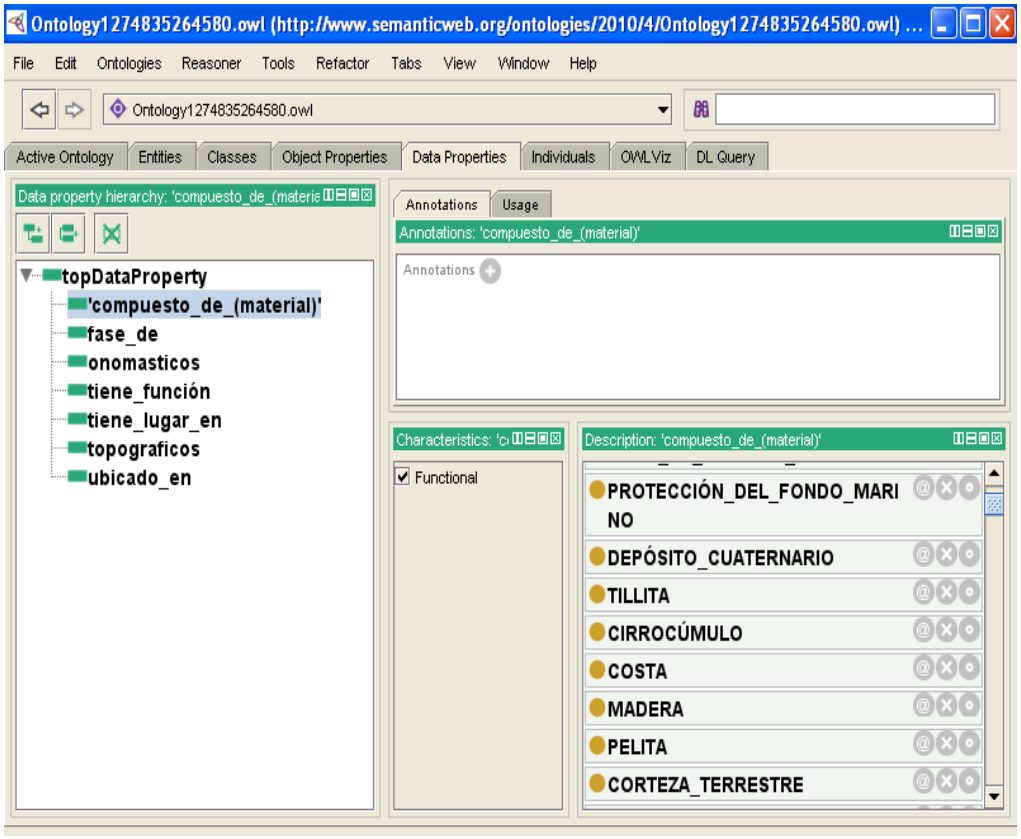



Figura. 87. Propiedades de los Objetos

### 6.3.1.4. – Anotación

En este segmento se trata de identificar las herramientas que ha de servir de base de anotación para los recursos asociados a la ontología.

Para Beltrán (n.d.) Las anotaciones semánticas, involucran el proceso de análisis, extracción y marcado de la información para enriquecerla semánticamente. Específicamente se trata de formular asociaciones específicas entre los conceptos y las cosas para luego insertarlos en una ontología (mapping), en la cual están definidas clases, atributos y relaciones que de conjunto con la anotación semántica facilitan el desarrollo operativo de

la Web semántica. Esta organización de contenidos debe ser leída mediante un lenguaje que facilite la lectura y transmisión de meta datos, para ello se ha creado un formato estándar llamado RDF, finalmente la anotación se almacena (Ver figura 88).

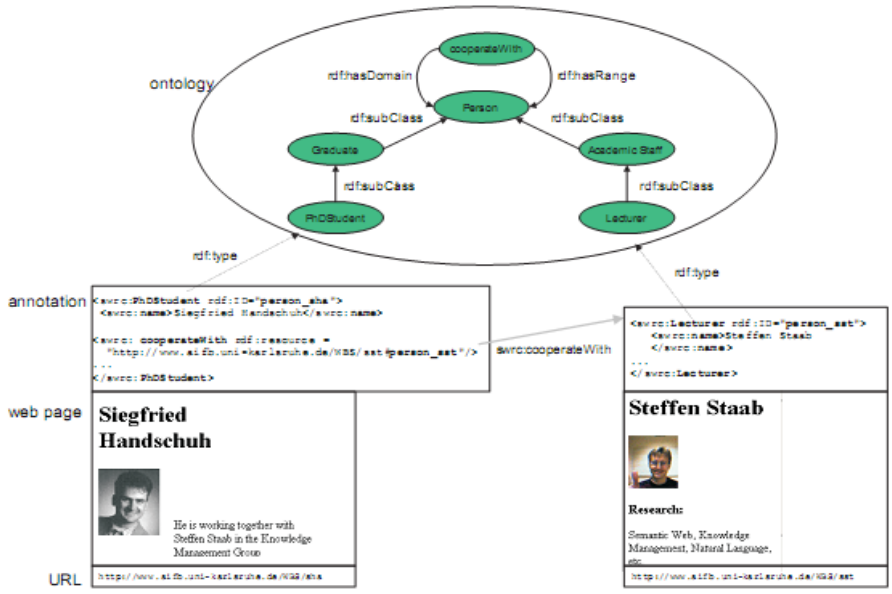



Figura: 88. Anotación (Handschuh and Staab, 2002)

### 6.3.1.4.1.- Clasificación de las herramientas de anotación

El proceso de anotación semántica se ha desarrollado rápidamente en los últimos años. Si bien en sus inicios este proceso era desarrollado con el concurso de instrumentos sencillos que servían de ayuda en la marcación semántica, es indiscutible que hoy en día este proceso se ha automatizado. Existen muchas formas de clasificar las herramientas para el marcado de información. En Senso (2008) se describe este proceso como anotación externa e interna, en Bernal se clasifican las herramientas de acuerdo al nivel de apoyo a la automatización y al método que emplean las mismas, la clasificación que describe (Beltrán, 2007) es la siguiente:

- **Anotaciones directas o manuales:** Son anotaciones hechas por el propio usuario para ello utilizan la herramienta RDF, estas se aplican cuando el contenido semántico es nulo**.**

- **Anotaciones manuales sobre contenido existente:** Se involucran en este apartado aquellas formas de anotación en las que el usuario no tiene nexos con el contenido que se le presenta. Esta anotación se hace con herramientas generales, donde el domino no es imprescindible para su efectividad.
- **Anotaciones automáticas mediante Reglas estáticas**: En esta categoría se encuentran aquellas herramientas cuya profundidad de trabajo permite algo más que un marcado convencional.
- **Anotaciones manuales durante la creación**: Estas anotaciones son netamente automáticas y facilitan el proceso de extracción y marcado de la información. Son efectivas si se aplican a un dominio específico. Las herramientas que se desarrollan para estas aplicaciones basan su acción en heurísticas, lo que indica que el análisis de domino es la herramienta que facilita el desarrollo de estos instrumentos.
- **Anotaciones basadas en aprendizaje**: Herramientas basadas en aprendizaje de reglas. Beltrán (Beltrán, 2007) sostiene que en esta categoría se encuentran las soluciones que se basan en el aprendizaje de reglas para realizar de forma autónoma la tarea de extracción de entidades semánticas del contenido, a diferencia del caso anterior, en que las reglas eran fijadas por el creador, estas aproximaciones permiten que el sistema incorpore reglas a partir de las ejecuciones de los usuarios, es decir, obteniendo la frecuencia de los procesos asumidos se establecen normas de aprendizaje.

En esta tesis se ha pensado en la clasificación observada en Senso (2008), donde se pone de manifiesto una clasificación más general que divide las herramientas en dos grupos: Herramientas de Anotación Externa y Herramientas de Anotación Externa.

#### 6.3.1.4.1.2.- Herramientas de anotación Interna

Los instrumentos y modelos para la anotación interna son aquellos que se centran dentro del sistema y que realizan automáticamente la anotación de los recursos sin el concurso del usuario del sistema. En esta categoría se

encuentran: Annotea, Smore, Yawas, MELITA, GATE, Briefing Asóciate, SemanticWord, Semantic Markup Plug-In for MS Internet Explorer, OntoMat Annotizer, KIM Semantic Annotation Platform, MnM, The SHOE Knowledge Annotator, AeroDAML, Trellis Web, las mismas se explican a continuación.

#### 6.3.1.4.1.2.1.-Annotea

Annotea se utiliza para anotar recursos que estén en consonancia con los estándares de W3C. Una de las propiedades especiales de Annotea es su flexibilidad para consultar documentos cuya estructura sintáctica sea el RDF. Annotea es parte de la Iniciativa para desarrollar la Web Semántica. Las anotaciones en Annotea se almacenan en los servidores de anotación como metadatos individuales, presentando al usuario una estructura de datos capaz de traducir estos metadatos e interactuar con un servidor de anotación con protocolo http. Según Beltrán y Senso (Beltrán, 2007, Senso, 2008) Annotea funciona de la siguiente forma: La operación general comienza cuando el usuario encuentra un documento, luego elige la sección del documento sobre la cual realizara la anotación, para lo cual el browser ofrece la posibilidad de ingresar la anotación, cuando el usuario ha incluido la anotación, el browser la marca con metadatos utilizando RDF, luego esta información es enviada al servidor, donde se almacena en una base datos, en la que queda pública.

#### 6.3.1.4.1.2.2.- Smore

SMORE es una herramienta que permite a los usuarios construir marcas en sus documentos en RDF. Para ello Smore utiliza ontologías que poseen términos o conceptos específicos de determinados grupos de usuarios. El objetivo de este software es la siguiente:

- Proporcionar al usuario un entorno flexible en el que pueda crear su página web sin demasiados obstáculos
- La construcción de marcas para permitir al usuario marcar su documento con un mínimo conocimiento de RDF y su sintaxis.

Sin embargo, el usuario debe ser capaz de clasificar semánticamente su conjunto de datos de anotación, es decir, debe tener habilidades suficientes para estructurar frases básicas en la semántica y la sintaxis de los documentos, además de proporcionar una referencia a las ontologías existentes en Internet a fin de utilizar las referencias más precisas en su propia página web / texto. Con esta herramienta el usuario también puede crear su propia ontología a partir de cero y tomar en préstamo términos de las ontologías.

#### 6.3.1.4.1.2.3.- Yawas

Es una herramienta para el desarrollo de anotaciones en Internet cuya eficiencia ha sido probada en diversos sistemas como Wikypedia. Su autor es Laurent Denoue. Las prestaciones de Yawas son demasiado escuetas y proporcionan un mediano rendimiento en los sistemas de estructuración del conocimiento.

#### 6.3.1.4.1.2.4.- MELITA

Melita es una ontología construida a partir de una herramienta de anotación de texto. Aplica una metodología particular de anotación con la intención de gestionar todo el proceso de anotación (Ver figura 89). Las dificultades que implica la anotación manual en otros sistemas pueden ser fácilmente automatizadas y manipulados todas por esta herramienta. Las principales competencias de Melita se pueden resumir en cuatro grupos:

- la tarea de gestión
- la tarea de extracción
- el aprendizaje
- la información de marcado en forma autónoma.

Esto se realiza gracias a la utilización de una interfaz inteligente junto con un potente algoritmo de extracción de información.

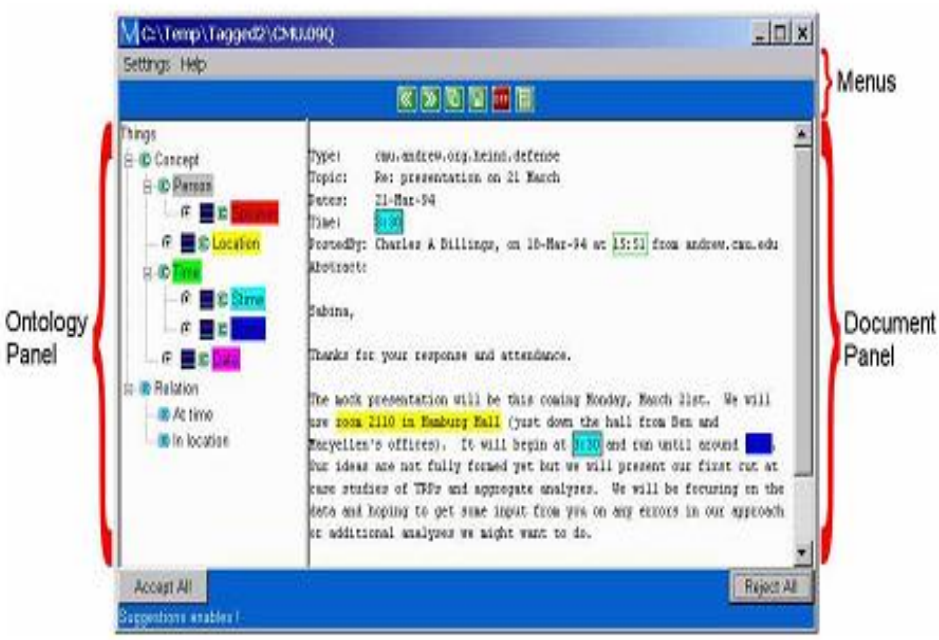



Figura 89. Melita (Dingli, 2003)

### 6.3.1.4.1.2.5.- GATE

GATE es una herramienta para realizar experimentos que implican la transformación del lenguaje humano. Entre sus clientes se encuentran las empresas de desarrollo de aplicaciones con componentes de procesamiento del lenguaje; profesores y alumnos de los cursos de lingüística y de lenguajes de computación. GATE comprende una arquitectura (SDK) de desarrollo y medio ambiente, y ha estado en desarrollo desde 1995 en el grupo de Sheffield NLP. El sistema se ha utilizado para muchos proyectos de procesamiento del lenguaje, en particular para la Extracción de Información en varios idiomas. GATE es financiado por el EPSRC y la UE (Unión Europea). Las potencialidades Gate están en el desarrollo de anotaciones usadas específicamente para el procesamiento y anotación de grandes léxicos. Su espectro de aplicación está concebido para el entorno europeo y para empresas medianas.

#### 6.3.1.4.1.2.6.- Briefing Associate

Un entorno basado en MS PowerPoint para la autoría de información semántica. El BA expone la información de la semántica a las descripciones de módulos externos llamados analizadores, que llevan a cabo servicios especializados o análisis del autor. Estos análisis facilitan información al autor de las anotaciones para que puedan ampliar o modificar la información, o producir documentos externos derivados de las descripciones.

#### 6.3.1.4.1.2.7.- SemanticWord

Un entorno basado en MS Word que integra tanto al documento como la anotación semántica de autor. SemanticWord es un conjunto de herramientas que faciliten la generación de contenido y de marcas. Es un sistema de anotación que construye anotaciones para ser reutilizadas cuando el contenido se necesite modificar. SemanticWord es una biblioteca personalizable de plantillas que contienen parcialmente el texto anotado por un experto.

#### 6.3.1.4.1.2.8.- Semantic Markup Plug-In for MS Internet Explorer

Semántica de marcas plug-in para MS Internet Explorer. Esta herramienta simplifica la autoría de marcado semántico de páginas Web a través de una interfaz funcional e inteligente, permitiendo la construcción de marcas a partir de protocolos de intercambio definidos con anterioridad por los usuarios.

#### 6.3.1.4.1.2.9.- OntoMa Annotizer

OntoMat-Annotizer es una herramienta interactiva de anotación. Apoya al usuario a crear y mantener una ontología basada en OWL, es decir, ayuda en la creación de márgenes de OWL, casos, atributos y relaciones. Incluye un navegador de ontología para la exploración de la ontología y las instancias, además de un navegador HTML que anota partes del texto. Es una herramienta basada en Java y proporciona una interfaz plug-in para diversas extensiones. El usuario hace función de anotador individual. Las personas que

quieren enriquecer sus páginas Web con OWL meta-datos pueden usar este sistema en lugar de anotar manualmente la página con un editor de texto. Se apoya en meta-datos. Está previsto que una futura versión incluirá un plug-in de extracción de información, que ofrezca un asistente que sugiera que partes del texto son pertinentes para la anotación. Este aspecto contribuirá a facilitar el tiempo y las tareas de anotación.

#### 6.3.1.4.1.2.10.- KIM Semantic Annotation Platform

KIM proporciona una efectiva gestión de conocimientos e información, además facilita la infraestructura y los servicios para la anotación semántica, automática, la indexación y la recuperación de la información en textos no estructurados y semi-estructurados. Dentro del proceso de anotación, Kim también realiza ontologías. Como una línea de base, Kim analiza los textos y reconoce las referencias a entidades (como personas, organizaciones, lugares, fechas). Intenta hacer coincidir la referencia con una entidad conocida, con una única URI y en la descripción en la base de conocimientos. Este proceso, así como el resultado de él, son las ofertas de KIM para la anotación semántica. Los metadatos resultantes de Kin se utilizan más adelante para la indexación semántica, la recuperación, visualización, y la construcción de los enlaces de los documentos. KIM es una plataforma que ofrece un servidor, interfaz de usuario Web utilizando Internet Explorer. KIM está equipada con una ontología de alto nivel denominada (KIMO), de alrededor de 250 clases y 100 propiedades. En cuanto a la tecnología subyacente, Kim está utilizando GATE, Sesame, y Lucene. Además esta herramienta posee, una base de conocimientos (KB KIM), pre-poblado con un máximo de 200 000 descripciones de entidad.

#### 6.3.1.4.1.2.11.- MnM

Es una herramienta que proporciona la automatización y semi-automatización de los procesos para anotar las páginas Web con contenido semántico. MnM tiene entre sus componentes un navegador Web con un editor de ontología. Proporciona APIs abiertas para enlazar con los servidores de la ontología y

para la integración de herramientas de extracción de información. Posee un arsenal lingüístico bien estructurado que le da fuerza para trabajar el inglés y el italiano, aunque su estructura esta resanada para varios idiomas. Funciona a través de un Server de ontologías. Posibilita y trabajar sobre documentos no estructurados y HTML.

#### 6.3.1.4.1.2.12. - The SHOE Knowledge Annotator

Parallel Understanding Systems Group Department of Computer Science University of Maryland .Es un producto de la Universidad de Maryland. Es un programa en Java que permite a los usuarios la marcación de páginas Web sin tener que preocuparse por el código HTML. Este software trabaja con ontologías específicas donde es necesario especificar el marco de referencia. Para utilizar esta herramienta se debe utilizar al menos una ontología o más de una si es necesario. Permite añadir, editar y eliminar botones. Con SHOE es posible anotar muchos documentos que están en el mismo dominio. El usuario elige una ontología y decide en que clase va a marcar en la misma y en ella hace las anotaciones pertinentes. Es un producto cuyo eclecticismo es muy efectivo para diversos usuarios.

#### 6.3.1.4.1.2.13.- AeroDAML

Se basa en técnicas de extracción desarrolladas a partir del PLN para asignar valores a clases y a contenidos. Se basa en ontologías que no tienen que ser especificas de un dominio. El usuario declara la URL o la ubicación del documento Web y la herramienta crea la anotación de forma automática. Esta herramienta permite trabajar con las concepciones de trabajo de DAML (DARPA Agent Markup Language), que es, una variante de organización del conocimiento. DAML también trabaja con RDF. Su principal aporte está en que su forma de accionar se basa en técnicas de PLN.

#### 6.3.1.4.1.2.14.- Trellis Web

Permite a los usuarios añadir sus observaciones, puntos de vista y conclusiones sobre la información que analizan, realizando anotaciones

semánticas a los documentos y a otros recursos on-line (Beltrán, 2007, Senso, 2008, Donés Rojas and Ortiz Rodríguez, 2006). Funciona como un modelo de análisis de dominio, pues estas estrategias de anotación declaradas por el usuario se utilizan en la mejora del sistema. Los operadores de este sistema necesitan un navegador Web capaz de soportar "frames". El mismo permite realizar anotaciones y ver las notaciones realizadas por el resto de usuarios. (Ver Figura).

### *6.3.2.- Herramientas de Anotación Externa*

Son instrumentos en lo que los procesos de anotación son realizados por el hombre ce conjunto con el software se denominan herramienta de anotación externa. A esta categoría pertenecen FRAMENET, Thresher, On Deep Annotation.

#### 6.3.2.1. - S-CREAM — Semi-automatic CREAtion of Metadata

Es un proyecto de anotación de páginas que facilita el desarrollo de estrategias cognitivas en lenguaje natural. Es un sistema de anotación basado en las herramientas de autor. Utiliza un sistema de administración de información, una meta-ontología, y un sistema de exacción y de reconocimiento. Todos los componentes del sistema tienen pluguins que intervienen en la extracción y anotación. Esta herramienta permite hacer marcaciones también de forma externa a las páginas Web (ver figura 90), siendo esta una de las cualidades que la distinguen como una herramienta multipropósito.

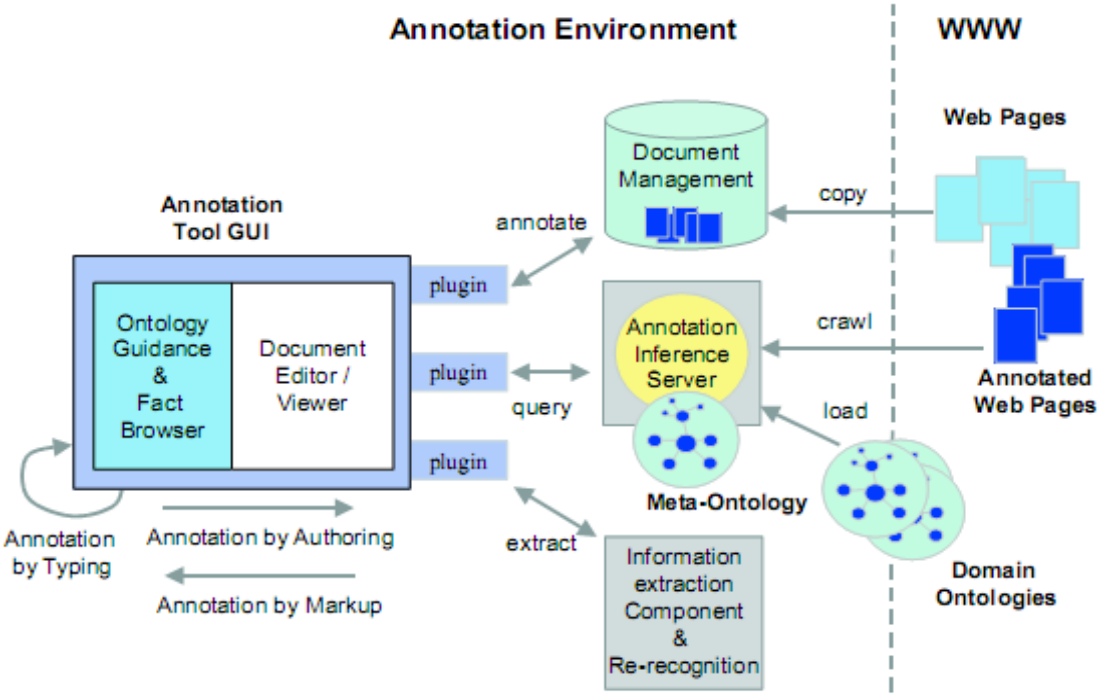



Figura: 90. Scream Anotación (Handschuh and Staab, 2002)

### 6.3.2.2.- FRAMENET

Es un proyecto de anotación semántica en FRAMENET en Español que tiene como objetivo crear una herramienta que facilite crear en los documentos anotaciones semánticas conceptuales e identificar marcas semánticas que identifiquen los predicados del español., además de anotar los predicados que se corresponden con los diferentes, marcos (Donés Rojas and Ortiz Rodríguez, 2006). Este proyecto se ampara en la teoría de la semántica de Marcos, usándola como herramienta teórica para desarrollar sus postulados. Permite la etiquetación, la consulta Web y la identificación de esquemas semánticas.

### 6.3.2.3.- Thresher

Según Handschuh y Staab (Handschuh and Staab, 2002) esta herramienta es una aplicación realizada en conjunto entre MIT y Google, su estrategia de construcción está basada en wrappers o modelos semánticos de contenido a partir de modelos de aprendizaje, generados mediante el desarrollo de Tree-Edit-Distance, o sea mediante técnicas de agrupamiento y clasificación de automática esta herramienta es capaz de generar distancias entre varias

entidades. El sistema tiene además un browser capaz de servir de mecanismos de hojeo de páginas y anotaciones, el cual está basado en la aplicación de la Tree-Edit-Distance. Además se presenta incrustado en un browser, lo cual le permite la relación con el usuario y los contenidos.

#### 6.3.2.4.- On Deep Annotation

Es un sistema de anotación de documentos basado en una ontología que permite obtener información a partir de un sitio Web. Tiene un servidor donde ocurren las marcaciones, mediante la relación cliente servidor, luego esa anotación es llevada a la ontología y se construye un mapa de reglas semánticas que son almacenadas en una base de datos. Este proceso permite que el usuario realice anotaciones en las clases y en las instancias de clases, lo que es considerado como una anotación profunda.

### *6.3.4.- Reflexiones en torno a la anotación Semántica*

Mucho se ha avanzado en campo de las marcas semánticas, sin embargo las herramientas que hoy en día tenemos en como sustento tecnológico para el desarrollo de estos procesos siguen siendo monodimensionales y cuando mas tridimensionales, esto significa que sus formas de construcción responden a patrones anquilosados, muy lejanos de la necesidades de los usuarios y de los proyectos de investigación actual. Es indiscutible el desarrollo de las soluciones automáticas, sin embargo en estas el marcado es restricto a documentos donde la semántica es una carga de análisis explicito. Lo que se observa en las aplicaciones estudiadas es la falta de herramientas descriptivas para otros objetos con condiciones semánticas diferentes, o sea para tratar información con otro tipo de estilo u otras formas de texto. Esto hace que se formule la siguiente interrogante ¿Cómo anotar información sonora, audiovisual o icónica? Esta interrogante no puede ser respondida por las actuales herramientas, su marco de acción está anclado esencialmente en elementos denotativos. Según Beltrán (Beltrán, 2007) las soluciones automáticas ofrecen una buena variedad de formas para llevar a cabo el enriquecimiento semántico del contenido, liberando a los usuarios de la carga de analizar, extraer y marcar

el contenido" esta afirmación está alejada de la realidad documental y obedece a un sentido netamente tecnicista que no toma en cuenta la Ciencia de la Información, la Documentación y los paradigmas dominantes en la Representación de la Información y el Conocimiento y la Lingüística Documental. Las soluciones autónomas poseen gran fortaleza si se quieren desarrollar un marcaje individual, pero el proceso se hace más lento pues depende de un entrenamiento del usuario para la realización del proceso. Según Beltrán (Beltrán, 2007) y Leiva (Leiva, 2008) una característica repetida en varias de las soluciones revisadas es que propenden por marcos de trabajo o suites con muchas funcionalidades para la extracción y el marcado, y con ontologías propias, incluso incluyen bases de conocimiento, además buscan presentar una interfaz amigable, y prefieren integración con clientes Web. Otro problema observado en las herramientas que se listaron anteriormente es que la mayoría de las herramientas comerciales utilizadas en los sistemas se basa en la anotación de HTML, lo que impide que los documentos generados en otras formas de texto puedan ser anotados.

### *6.3.5.- Anotaciones Semánticas en Ontosatcol*

Las anotaciones semánticas en esta aplicación se han desarrollado utilizando un esquema de metadatos generados a partir del Dublín Core (DC).

DC es una estructura de metadatos desarrollado por el Grupo de Dublín (Dublín, Ohío, Estados Unidos) , nombre de la ciudad donde se creó este estándar y que generó la organización DCMI (Dublin Core Metadata Initiative), dedicada a la promoción y al desarrollo de un estándar de metadatos necesarios para la descripción y organización del conocimiento electrónico. Este modelo de metadatos se sustenta en vocabularios especializados de metadatos cuya sintaxis facilita la interoperabilidad de los sistemas y la edificación de sistemas basados en inteligencia artificial para el descubrimiento de conocimiento.

Cuando se implementa un sistema en Dublin Core se utilizan etiquetas en XML, basadas en RDF Resource Description Frameworkn (Ver Figura 91 - 92).

Debido a que este esquema de metadatos no posee las condiciones para la descripción exhaustiva de diversos recursos en red, entre los que se encuentran fotos, videos, etc. Los elementos que se usaron en la descripción se basan en la estructura clásica del Dublín Core y se apoyan en los enunciados teóricos de (Moreiro, 2004, Moreiro, 2006, Moreiro, 1996). Los elementos utilizados aquí son los siguientes:

- **Comment:** Describe la definición del término
- **Contributor:** Aquellas personas que son colaboradores del documento, pero no autores directos.
- **Creator:** Autor del documento
- **Date:** fecha y año.
- **Country:** Lugar.
- **Description:** Usado para describir datos físicos en el documento.
- **Indentifier:** Indica la URL de un sitio Web
- **Languaje:** Definido para declarar el idioma en que está el documento.
- **SeeAlso:** Permite declarar palabras claves no oficializadas y términos en inglés, ampliando el ámbito de búsqueda de información.
- **Subjet:** Permite indizar los documentos a partir de un vocabulario especializado construido a tales efectos. En subjet encontramos los siguientes sub-elemetos:

    - **Onomástico:** indiza documentos que tengan una persona o un grupo de personas en su contenido. Amplía las posibilidades de búsqueda de los documentos. Se usa con documentos cuyo contenido tenga niveles audiovisuales.
    - **Connotación:** Muchas documentos connotan situaciones y problemas, por tanto es necesario declararlos en la descripción de documentos audiovisuales.
    - **Denotación:** Se usa para declarar aquellos elementos que son denotados en una imagen o un material audiovisual.
    - **Topográfico:** Sitúa el fenómeno en un espacio geográfico.
    - **Title:** Título del documentos

- **Type:** Tipo de Recurso Web
- **Tipo de artículo:** Enuncia el nivel de especialización del documento.

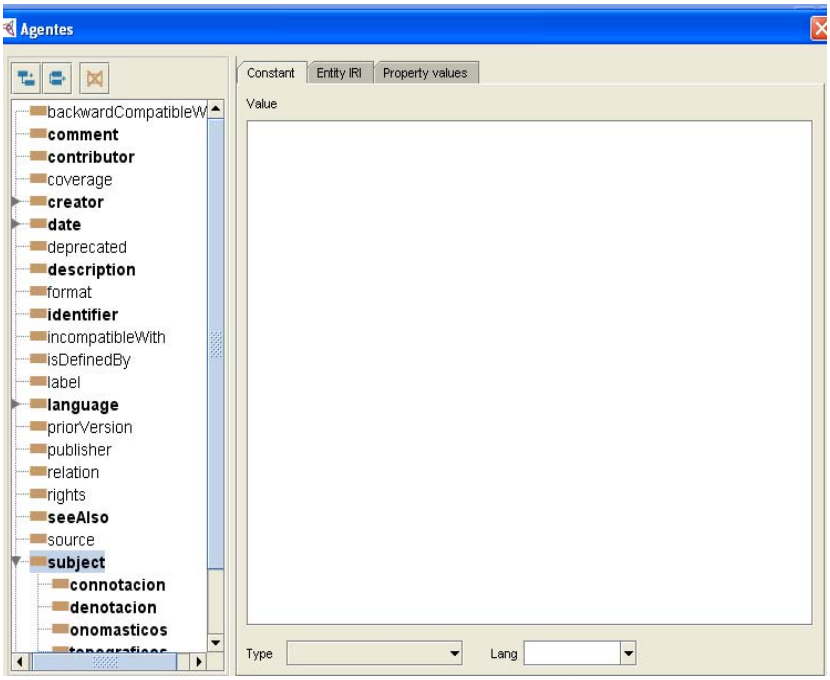



Figura: 91 Anotaciones en Puertoterm

**Algunos Elementos del DC**

```
<owl:Class
rdf:about="http://www.semanticweb.org/ontologies/2010/4/Ontology1274835264580.owl#AGUA">
   <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/2010/4/Ontology1274835264580.owl#Agentes_Naturales"/>
   <dc:subject>BOSQUES</dc:subject>
   <dc:subject>RIOS</dc:subject>
   <dc:title>Laguna Mata Redonda
</dc:title>
   <dc:subject>PROTECCION DEL CLIMA</dc:subject>
   <dc:title>Agua, recurso estrat&#233;gico garantizado en el Estado de M&#233;xico
</dc:title>
   <rdfs:seeAlso>LAGUNA</rdfs:seeAlso>
   <dc:title>Aguas salvajes y de arroyada</dc:title>
```

Figura 92. Elementos de Dublín Core

Finalmente se han de declarar las instancias de la ontología que están conectadas a los siguientes elementos:

1. Términos Sinónimos
2. Términos Homónimos
3. Términos Merónimos
4. Términos Hiperónimos
5. Documentos
6. Verbos
7. Sustantivos
8. Pronombres
9. Pronombres Anafóricos
10. Imágenes y Fotografías

### *6.4. Agentes*

La arquitectura principal presentada en esta tesis se basa en el trabajo coordinado y colaborativo de diversos agentes, encargados de las tareas de minería, indización, clustering, resumen, etc. Aunque parece que es un concepto relativamente aceptado por la mayoría de expertos, es complejo y confuso trabajar con una definición clara sobre qué es un agente. Posiblemente el origen de esta confusión se encuentre, por un lado, en el intento de establecer una definición única y universalmente aceptada –como ocurre con otras muchas-. Por otra parte, como explica Foner (Foner, 1993), el hecho de que el trabajo con agentes capte la atención de una gran cantidad de especialistas de diferentes áreas de conocimiento (principalmente la psicología, la sociología, la ingeniería del software, la inteligencia artificial y la documentación) provoca que cada uno lleve la definición a su terreno, centrándose únicamente en la óptica de su campo de aplicación. Es algo muy parecido a lo que sucede con las ontologías.

Existen una gran cantidad de definiciones más o menos aceptadas dentro de la comunidad científica (Foner, 1996, Franklin and Graesser, 1996). Por lo

general, todos ellos lo caracterizan como un sistema informático, ubicado dentro de un entorno concreto que es el que les permite realizar determinadas acciones de manera más o menos autónoma y flexible. Para ello, en muchas ocasiones deben ser capaces de cambiar su entorno (Russell and Norvig, 2003).

Otros autores defienden la teoría de que el agente debe contar con dos habilidades, a priori ortogonales, con son la capacidad de ejecutar acciones de forma autónoma y la de llevar a cabo razonamientos orientados al dominio en el cual está trabajando (Virdhagriswaran, 1994). Una parte de esta filosofía es compartida por otros autores que la complementan, observando que el entorno debe ser dinámico y complejo, lo que permite que el agente pueda realizar de manera más eficaz las tareas para las que ha sido diseñado (Maes, 1994).

Para otros autores es más importante incidir en la capacidad persistente de estos programas. Dicha persistencia permite diferenciar a los agentes de las simples subrutinas de programación. El concepto de la persistencia implica que los agentes puedan tener ideas propias sobre cuál es el mejor método para ejecutar tareas, priorizar entradas en agendas, etc. (Smith et al., 1994).

Todo esto concluye en la idea de que los agentes realizan constantemente tres funciones: percibir condiciones dinámicas en el entorno, actuar afectando a las condiciones del entorno para el que trabajan y razonar para interpretar lo percibido. Es decir, que son capaces de resolver problemas por medio de acciones que también determinan ellos (Hayes-Roth et al., 1995).

Para el autor de este texto, los conceptos son hijos de su tiempo y de sus aplicaciones, por lo que la definición de agente que vamos a tomar está basada en la propuesta desarrollada por Leiva (Leiva et al., 2009), donde se define un agente como una máquina software para realizar estrategias cognitivas, dimensión más que aceptada dentro del terreno de la Documentación.

### 6.4.1.- Diseño de los agentes

En general, el sistema de información que se encargará de realizar los procesos automatizados cuenta con 6 agentes con sus respectivos autómatas finitos. Este tipo de autómatas se caracterizan por permitir formar un conjunto

finito de estados y otro de transiciones de estado a estado que se pueden dar sobre símbolos de entrada tomados a partir de un alfabeto ∑. Cada símbolo de entrada cuenta con una transición a partir de casa estado que, en la mayoría de los casos, denota el comienzo del agente (de ahí que se denomine inicial).

Básicamente, los autómatas finitos pueden ser deterministicos y no deterministicos (hay más tipos, pero su descripción se alejan de los objetivos del trabajo). Los primeros permiten trabajar mejor con estados fijos de comportamiento, mientras que los segundos ofrecen más opciones para la transición y aceptan varios modos iniciales para el comienzo del agente. Precisamente por estas opciones que ofrecen se optó por este tipo de autómatas, definidos como $K * \sum U \to P K, q_0 \in K, F \subseteq K$, donde K es el conjunto de estados, ∑ el alfabeto, δ la función de transición, F el conjunto de estados finales y qo el estado inicial.

Para llevar a cabo el trabajo con texto se han analizado cuales serían las operaciones esenciales para la transformación del corpus en cadenas de caracteres de orden binario. Una vez determinadas dichas operaciones se realizó la construcción de los autómatas, siguiendo la metodología propuesta por Aguirre y Arroyo (Arroyo et al., 2008), que cuenta con las siguientes fases:

**Etapa 1: obtención de los estados.** Para ello se obtuvo un modelo de estrategias cognitivas basadas en el estudio de 12 resumidores, el resultado fue un modelo que describe las acciones cognitivas siguientes: lectura, normalización, búsqueda de información y determinación de herramientas de relevancia.

**Etapa 2: determinar los nodos.** Establecer los nodos que son necesarios para cada uno de los procesos que se están construyendo. Para ello se tuvo en cuenta que no existiesen repeticiones de procesos ni acciones que pudieran extrapolarse o solaparse unas con otras.

**Etapa 3: desarrollo de los estados.** En esta etapa se describen las transiciones que se declaran para el agente desarrollando la tabla de estados necesarios para el sistema. Los estados iniciales y los estados finales del sistema son modelados para cada uno de los agentes. Para ello se realizó una

herramienta de simulación de estados, Estadist 29 (Domínguez, 2011a), que declara de forma aproximada los estados invariables de cada agente (Ver tabla 51).

| Estado Actual | Estado Próximo | | Salida | |
|---|---|---|---|---|
| | I0 | | Z0 | Z1 |
| | 0 | 1 | | |
| I | II | IV | 0 | 1 |
| II | III | III | 1 | 1 |
| III | IV | IV | 0 | 1 |
| IV | III | V | 1 | 1 |
| V | IV | I | 1 | 0 |
| VI | V | 0 | 1 | 0 |

Tabla. 51. Estados de los agentes.

**Etapa 4**: **minimizar la tabla de estados y construcción de árboles.** Para ello se empleó el método propuesto por Huffman (Huffman, 1952), ya que tan sólo se necesitan los estados finales, no las transiciones (Tabla 52).

| | Asignaciones | |
|---|---|---|
| Estado | No.1 | No.2 |
| I | 0 | 0 |
| II | 12 | 12 |
| III | 11 | 11 |
| IV | 10 | 10 |
| V | 9 | 9 |
| VI | 9 | 9 |

Tabla. 52 Frecuencia de ocurrencia de los estados.

Para aplicar el modelo de Huffman se crearon varios árboles (uno por cada uno de los términos) en un nodo sin hijos (paso 1). Cada árbol se etiquetó con un símbolo y con la información de su frecuencia de aparición. A continuación se seleccionan los dos árboles de menor frecuencia y se unen, creando un nuevo árbol (paso 2). La etiqueta que empleará será el resultado de la suma de las frecuencias de las raíces de los dos árboles. El árbol resultante tendrá como hijos los dos árboles que lo generaron, y contará con dos ramas, que también

se etiquetarán (con un 0 la rama de la izquierda y con un 1 la de la derecha). Después se repite el paso 2 hasta que sólo quede un árbol.

**Etapa 5: diseño del agente.** De acuerdo a la simulación realizada por software, se establecen los tiempos del desarrollo de cada función y se calculan la complejidad del sistema. Esto facilitó el desarrollo de una aplicación de alto nivel.

### 6.4.2.- Trabajo con el corpus

Como se comentó anteriormente, los agentes generados hasta este momento trabajan contra el corpus empleado en el proyecto de investigación PuertoTerm (Senso, 2007), especializado en ingeniería de puertos y costas, aprovechando las entradas paradigmáticas. Domínguez (Domínguez-Velasco, 2009) describe en su modelo un agente capaz de automodificarse. Esta mutación se realiza para minimizar los errores paradigmáticos que llevan consigo los textos impresos en el proceso de búsqueda.

Como el objetivo final es resumir, clasificar y marcar los términos del corpus como si fueran entradas de un diccionario, es necesario que las acciones descritas en el procesamiento de dicha información estén claramente representadas mediante técnicas de clasificación y de resumen. Otro elemento a tener en cuenta es el establecimiento de relaciones de equivalencia entre los términos sinónimos, construyendo índices de clasificación eficientes.

#### 6.4.2.1.- Estrategia de procesamiento del texto

Ante la inexistencia de analizadores morfoléxicos lo suficientemente potentes en lengua española que permitan la detección de unidades terminológicas y estructuras léxico-semánticas, se optó por etiquetar en xml los elementos estructurales de los textos del corpus, almacenando los mismos en ficheros. Dichos ficheros serán los que procesarán los agentes.

Tal y como defienden diferentes autores (López-Huertas, 2008, Faber and Mairal-Usón, 1999), se optó por el empleo de un lenguaje que facilitara el procesamiento de los elementos del texto por medio del etiquetado. Para ello se empleó un editor XML para el marcado, y una herramienta de análisis léxico, Wordsmith Tools Wordsmith Tools, distribuido por Oxford University Press y

que permite explotar grandes conjuntos de textos mediante búsquedas basadas en parámetros contextuales o estadísticos http://www.lexically.net/wordsmith/ . Dicha aplicación permitió inferir el conocimiento especializado en ingeniería de puertos y costas por medio de los términos más empleados y sus concordancias. La identificación de las palabras clave permitió el modelado conceptual de la ontología que articularía todo el conocimiento de la disciplina. Los lemas más frecuentes permiten identificar las categorías conceptuales sobre las que se fundamenta la definición de los términos del texto.

La ontología sobre la que se desarrolla la aplicación descrita en este trabajo está formada por una estrategia cognitiva y semántica donde se expresan las relaciones semánticas, interpersonales y algunos de los agentes que operan sobre ella.

#### 6.4.2.1.1- Agentes de PUERTOTEX: Funciones

En esta aplicación, un conjunto de agentes realizan determinadas funciones en el trabajo con el texto, en los procesos de resumen, búsqueda de información y normalización (Ver figura 93). A continuación se explican las funciones de cada uno de ellos:

- **Agente de Lectura:** Se encarga de realizar la primera estrategia cognitiva que hacen los resumidores que escriben para el dominio. Leen el documentos y comprueba que esté acorde con el dominio de la ingeniería de puertos y costas, para ello lee el texto marcado en XML y contrasta el conocimiento contra la ontología y la base léxica, determinando que el texto es relevante para el dominio ya que posee términos que son relevantes para él. Envía al agente de resumen el texto para que este lo resuma.
- **Agente de Resumen:** Este agente toma el texto que le envía el agente de lectura y realiza dos procesos. Primero extrae todas la oraciones candidatas para el resumen de acuerdo a las calificaciones declaradas en el análisis del discurso o sea hace un resumen por extracción y en un segundo paso hace un resumen por abstracción, es decir reescribe el texto utilizando reglas heurísticas que tiene implementadas, mediante

estas reglas logra la cohesión, la coherencia y el balance textual. En el proceso de extracción asigna pesos a las palabras que existen en la ontología de acuerdo a los valores que se establecen en las reglas PUERTOTERM y en las Reglas de Léxicas, para buscar la cohesión del texto, se auxilia de las Reglas PUERTOTERM de segunda Generación.

- **Agente de Desambiguación:** Luego de estar realizado el resumen este agente recibe el texto y reconoce en él, signos de puntuación erróneos y palabras fuera de contexto, verificando los vocablos contra la ontología.
- **Agente de Normalización:** Se encarga de determinar la estructura del resumen. Recibe el texto del agente de desambiguación y le asigna al texto el segmento de referencia del artículo, así como su URL, para que la presentación del resumen permita determinar dónde está el texto fuente y su corpus de referencia.
- **Agente de Búsqueda:** Es el agente de más trabajo en el sistema, hace búsquedas a través de la ontología y a través de la base de datos léxica desarrollada. En la ontología se encarga de buscar por autor, título, materia, tipo de documento, utilizando las bondades de Dublín Core, también permite buscar en redes sociales desarrolladas a partir de los datos de los autores que se encuentran en la estructura FOAF. FOAF facilita la búsqueda y recuperación de los siguientes elementos: institución del autor, título del proyecto, correo electrónico del autor y los autores que colaboran con él.
- **Agente de Redundancia:** Conecta la solicitud del usuario con los resúmenes que ya tiene guardado en la base de datos, y cuando se realiza una solicitud de un texto que no tiene le pide al usuario que indique donde está el texto para hacer el tratamiento del texto o si ya lo tiene se encarga de buscar el resumen y entregarle el texto al usuario

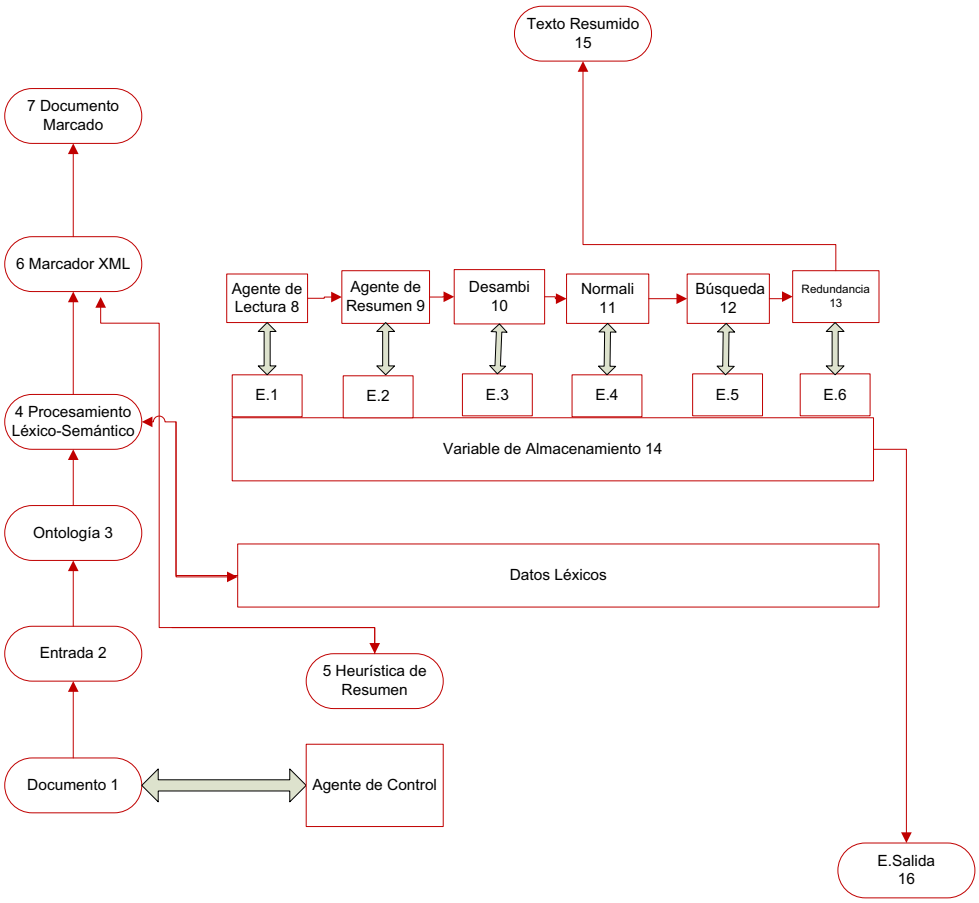



Figura. 93 Diagrama de agentes

Como se observa en la figura 91 cada documento entra al sistema y su contenido semántico es descargado en la ontología para ser recuperado, luego se realiza un procesamiento léxico-semántico a partir de la base de datos léxica del sistema. El siguiente paso consiste en realizar las marcas de discurso en el texto de acuerdo con la heurística (estrategia cognitiva) tomada de los resumidores humanos que participan en este estudio y da como resultado el documento marcado en XML. Luego de estar marcado el documento el texto pasa por al agente de lectura que se encarga de leer el documento al igual que un lector humano. Cuando este agente termina la lectura y comprueba que el texto está marcado correctamente, entonces envía el texto al agente de resumen, el que utiliza varias estrategias de resumen para determinar las oraciones relevantes, asignar pesos a los términos y extraer las oraciones (hace el resumen por abstracción). El resumen resultante de este paso aún no es legible, por tanto el agente de de desambiguación tiene que darle sentido a los términos utilizando la ontología. Después de la desambiguación se pasa a la normalización del texto, es decir, se colocan los

puntos, las comas y las unidades necesarias para que el texto sea legible. Resumido el texto se almacena y está listo para que el agente de búsqueda localice información (teniendo en cuenta los valores que posibilita la ontología) y además interrogue al agente de redundancia para cuando se realice una solicitud de información que ya ha sido procesada.

### *6.5.- Visualización*

Como todo sistema de Representación del Conocimiento se determinó desarrollar estrategias para la Recuperación y Visualización de la Información obtenida. Como la información que usa el sistema se almacena en diversas estructuras de conocimiento: bases de datos y ontología, se decidió realizar consultas (querying) y búsquedas analíticas (browsing), lo que facilitaría la integración de esta herramienta con otras utilizadas en la Universidad Central de las Villas Cuba, entre las que se destaca: Scientia y CENCOM. Otro elemento en que se decidió trabajar fue en la visualización de la información para que la representación conceptual fuera más visual, para este aspecto se crearon restricciones para cara tipo de relación con diversos niveles de cardinalidad en algunos casos, lo que facilita la comprensión de las clasificaciones construidas en la ontología. El uso de herramientas de visualización en PUERTOTEX continua la idea que se desarrolló en el primer sistema del grupo de investigación Puertoterm, dotando al sistema de un medio de representación de la información capaz de ser utilizado por otro sistema similar, facilitando la inferencia de conocimiento en el acto de búsqueda ya que dimensiona la estructura de un concepto.

La Visualización de la Información se hace muy compleja en estos sistemas, sobre todo si se pretende hacer esto desde Cuba. Lo primero que hay que tener en cuenta es que este sistema está desarrollado en Python, un lenguaje que aunque se utiliza hace ya algún tiempo obliga a que haya que hacer adaptaciones constantes ya que hay visualizadores que no se entienden con su sintaxis. Para trabajar con Python y Qt se revisaron algunas tecnologías que se utilizan para la visualización en la Web Semántica, específicamente proyectos y herramientas:

Flash, patrocinado por Adobe System, una herramienta de elevada popularidad en el Web. Facilita la construcción de animaciones vectoriales. Los gráficos que genera permiten construir animaciones bidimensionales de poco peso y fáciles de descargar. Flash se aplica en **MACE**[36] (Metadata for Architectural Contents in Europe) (Ver figura 94), Proyecto pan-europeo, especializado en información sobre arquitectura, el cual visualiza la información de sus múltiples taxonomías y bases de datos. Su sistema de búsqueda posee una interfaz dividida en dos ventanas. En la superior se encuentra la representación de las relaciones en Flash, y en la segunda surgen los documentos conectados con los conceptos expresados en los diferentes nodos.

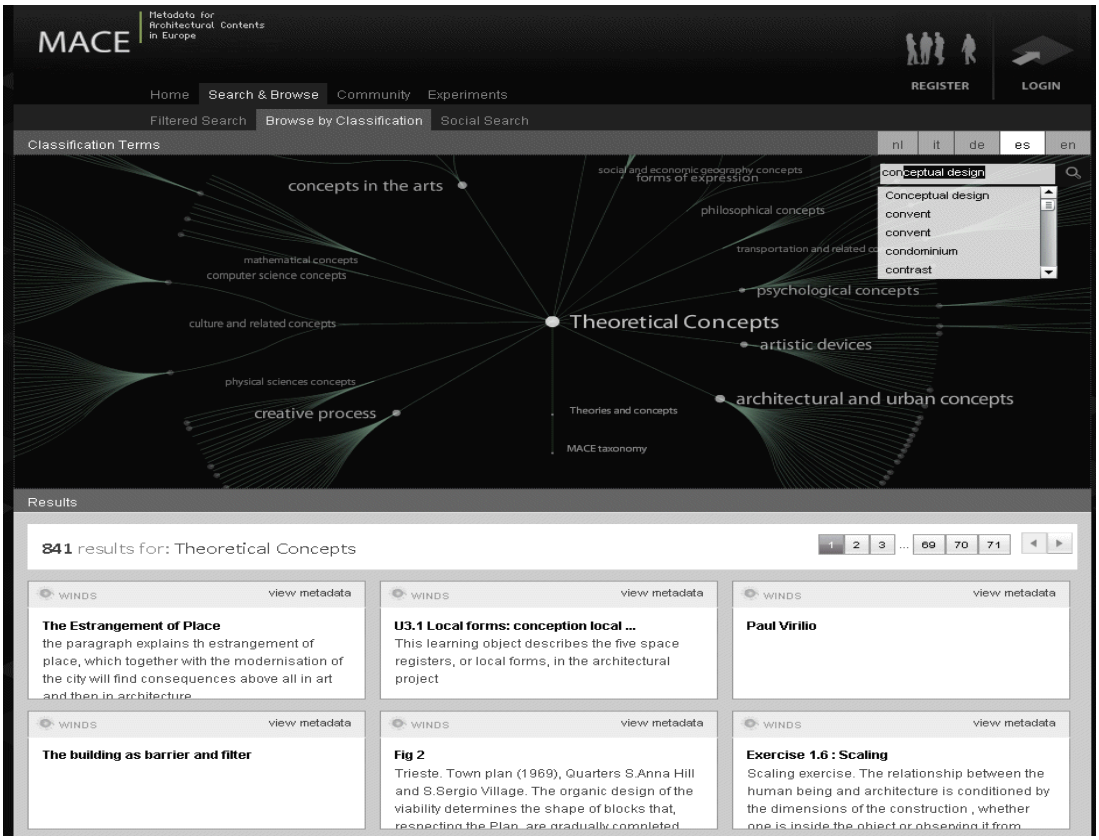



Figura 94. . Interfaz de consulta del proyecto MACE (Senso, 2009)

Otro proyecto que utiliza Flash es VisuWord, diccionario en línea especializado en lengua inglesa, capaz de generar mapas con las relaciones semánticas y agrupar los resultados de las consultas de acuerdo a su afinidad y tipo. Para lograr estas prestaciones se vale de una base de datos desarrollada por la

[36] http://portal.mace-project.eu/

Universidad de Princenton, una base de datos de carácter abierto (Wordnet[37]). Como afirma Senso (Senso, 2009) Es -en síntesis- un mashup de mapeo de palabras (Ver figura 95).

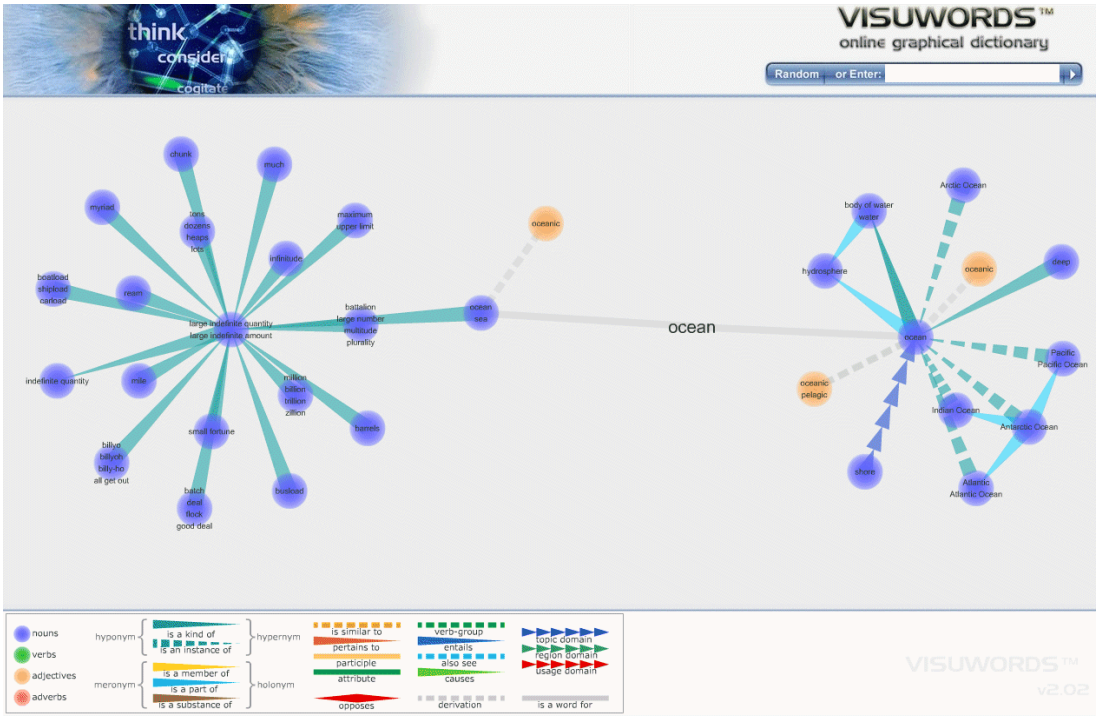



Figura 95. Interfaz de consulta de VisuWord (Senso, 2009)

Es indiscutible el atractivo visual que permiten estas herramientas, sin embargo no son compatibles con las tecnologías que utiliza nuestra herramienta PUERTOTEX. El problema principal estriba en la linealidad de las relaciones, pues solo es capaz de explicitar relaciones de índole taxonómica por lo que se deja fuera la posibilidad de representar relaciones cuya naturaleza conceptual sea más fuerte. Otro de las inconvenientes que genera es la imposibilidad de enlazar (dejando fuera la posibilidad de expresar relaciones más fuertes, además de no contemplar la posibilidad de unir) objetos externos (imágenes, vídeos o, simplemente, páginas Web), algo que en versiones futuras aparecerá en este software (Ver figura 96).

[37] http://wordnet.princeton.edu/

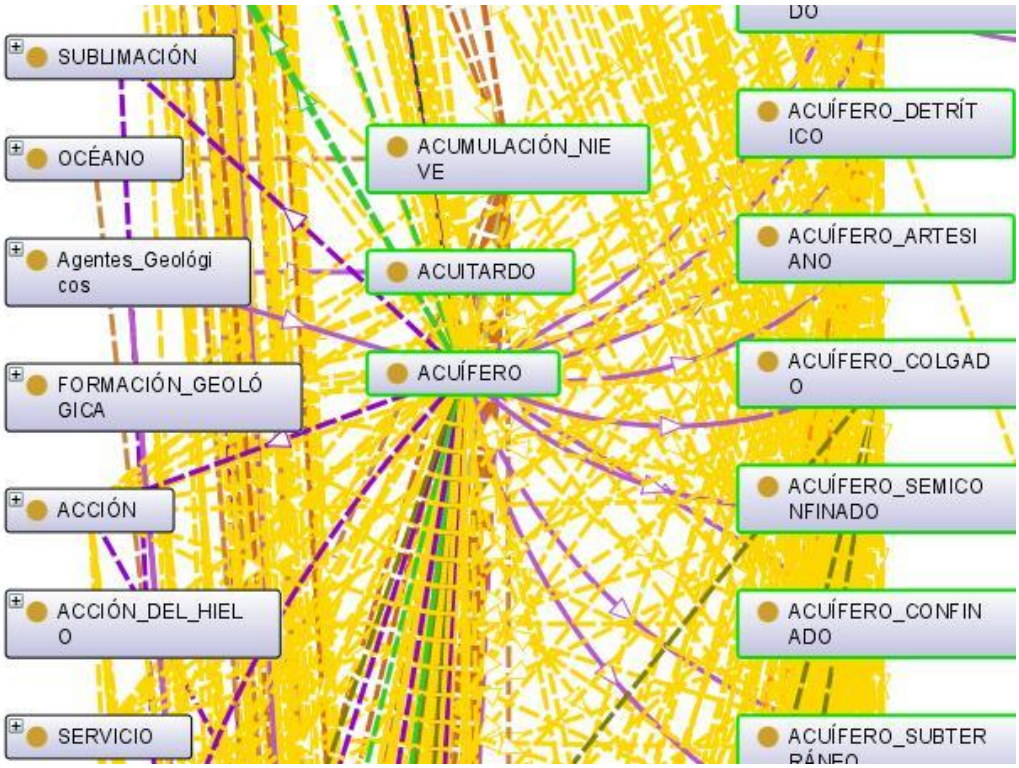



Figura. 96 Interfaz de Protégé con errores en la unión de clases

Otros problemas que generaba esta opción son los asociados al trabajo con Flash:

- La necesidad de plugin para ver las animaciones. Según Senso (Senso, 2009) en la actualidad muchas empresas restringen a los usuarios para que no puedan instalar programas en los equipos con el fin de que no ejecuten programas que puedan ser peligrosos para la red y/o ejecutar virus. Los navegadores por defecto vienen sin el plugin de Macromedia.
- Las animaciones no están anexadas directamente al navegador.
- Visualiza los textos con cierto grado de dificultad si este es pequeño.
- Realizar y mantener ficheros en Flash es caro y no permite que el usuario pueda usar otro sistema, ya que es un sistema propietario.

Como segunda opción también se ha visto SVG[38] (Scalable Vector Graphics), un plugin que puede manejar Qtdesigner y Pyqt4. SVG es un lenguaje sustentado en documentos formateados en XML que facilita la ceración de gráficos bidimensionales y animaciones. Posee varias versiones: SVG Basic,

[38] http://www.w3.org/Graphics/SVG/

SVG 1.1, SVG Print y SVG 1.2 o SVG Tin (para dispositivos móviles). Desde la óptica de este proyecto SVG posee una gran desventaja: la imposibilidad de rehacer mapas nuevos en SVG a partir de una selección realizada por un usuario cuando pincha en un nodo concreto del árbol.

Otra opción muy difundida en la visualización son los Applets de Java, dadas las posibilidades que ofrece este lenguaje de programación y sus nexos con Python y Qt y la optimización de sus componentes, sin embargo su uso haría que el sistema se hiciera muy pesado.

En la búsqueda de nuevas formas de visualización se ha podido ver que cuando estas herramientas se utilizan en bases de datos soslayan la información semántica, lo que las hace no viables para la construcción de la herramienta. Se reconoce que hubiese sido mucho más fácil desarrollar visualizaciones similares a las que poseen las ontologías, debido a su riqueza visual y expresiva, pero aún son muy restringidas al propio editor donde se desarrolla la ontología (Jambalaya, Knowledge Tree, Ontoviz, TGViz, etc.).

En nuestra búsqueda de información sobre visualización también estudiamos ThinkMap[39], que combina representaciones en forma arbórea (Kroeker, 2004). Este software ha sido empleado también en la versión que se realizó en España del Proyecto Puertoterm, así como en Visual Thesaurus[40], Sony Music Licensing[41] y la National Oceanic and Atmospheric Administration[42] de los EUA. ThinkMap facilita la organización de la información, ofreciéndole un sentido pragmático y semántico a las relaciones que visualiza.

Según Senso (Senso, 2008) el programa facilita una infraestructura (framework) que permite desarrollar aplicaciones muy flexibles -con un innovador entorno- extremadamente eficientes y fácilmente escalables y que puedan interactuar con diferentes fuentes de información. Estas prestaciones son relevantes para PUERTOTEX, ya que la información para la visualización podría alimentarse desde la ontología Ontosatcol.

[39] http://www.thinkmap.com
[40] http://www.visualthesaurus.com
[41] http://www.sonymusic.com/licensing
[42] http://www.noaa.gov/

ThinkMap utiliza un amplio cúmulo de algoritmos de agrupamiento que le permiten visualizar grandes cantidades de información con menos costo y eficiencia si se le compara con otros sistemas, lo que permite la construcción de relaciones dinámicas entre muchos nodos y la realización de búsquedas a partir de las consultas iniciales. En la figura 97 se muestran las diversas opciones de búsqueda, historial, y representación de la información implementadas en PuertoTerm a partir de esta herramienta. La única problemática de Think map es el costo de la aplicación más de 10.000 euros de licencia.

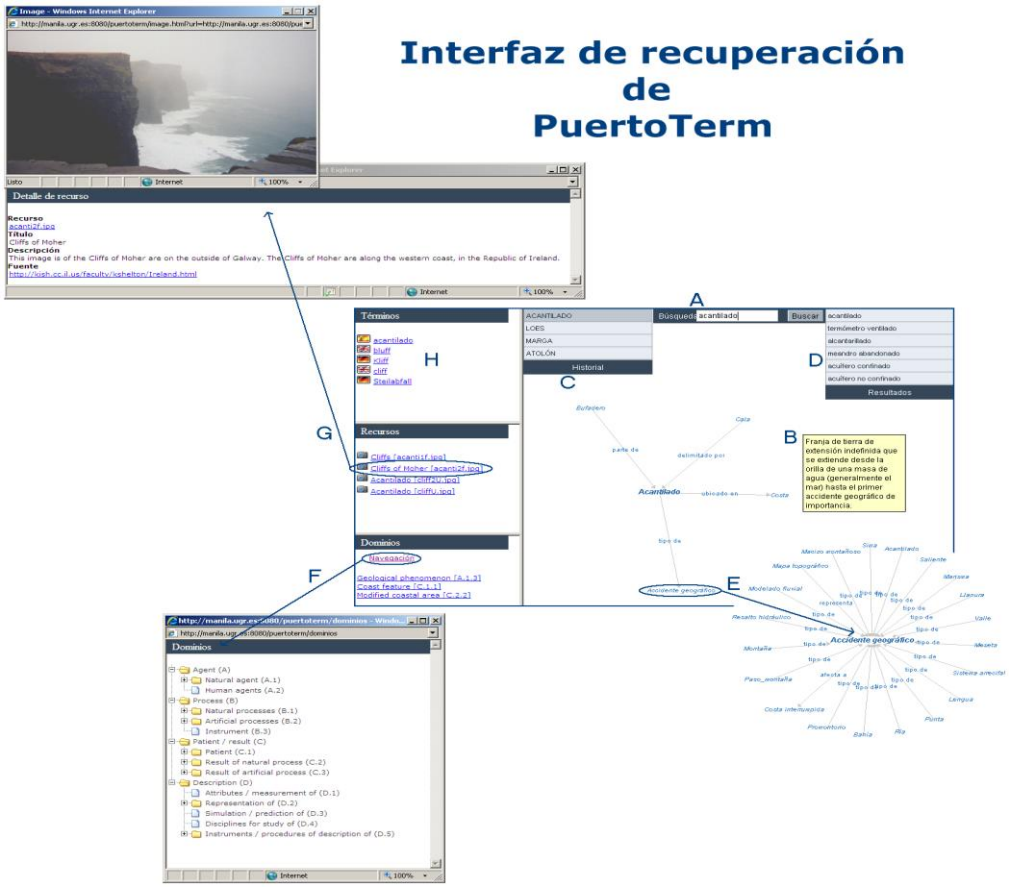



Figura 97. Aspecto de la interfaz de recuperación de la información de PuertoTerm. (Senso et al., 2007).

Otro aspecto de especial relevancia para nuestros propósitos radica en la facilidad que tiene el programa para integrar contenidos multimedia. Éste nuevo tipo de información visual complementa y amplia el contenido textual, como queda atestiguado en numerosos ejemplos (Le Grand Dictionnaire

Terminologique[43]) o la inclusión de opciones para la gestión de contenido multimedia en la mayoría de las herramientas para la gestión terminológicas.

Finalmente se estudiaron los Mindmaps, Concept Maps, y otra serie de representaciones de la información, en las que se apreció que solo responden a herramientas específicas. Los Concept Maps están orientados a visualizar las relaciones entre conceptos (Novak, 1991).

Observados todos estos casos se decidió utilizar QGraphicsView, una opción para construir gráficos y redes desde QT y Qtdesigner. Con esta opción es posible construir gráficos dinámicos capaces de autogenerarse y se obtienen visualizaciones de gran nivel. Unido a todo esto es importante destacar la portabilidad de este sistema que lo hace atractivo para los diversos usuarios de PUERTOTEX. El amplio rango de usuarios a los que PUERTOTEX brinda servicios está dividido fundamentalmente en tres grupos. El primero corresponde a los lingüistas documentalistas que alimentan y desarrollan la base de conocimientos (ontologías). El grupo de expertos en Ingeniería de Puertos y Costas conforman el segundo de ellos. Finalmente, los usuarios finales del sistema podrían abarcar personas que se desempeñan en la Gestión de Información hasta decisores que tengan que activar conocimiento sobre el tema.

Para lograr relaciones mucho más complejas que las denominadas (is-a) el autor concibió la creación de restricciones de cardinalidad (véase figura 98). Para ello el sistema posee una serie de constructores definidos para su uso en la construcción de relaciones entre los dominios y los rangos, con los cuales se logra una especificación de la lógica descriptiva del sistema:

- Cardinalidad Mínima: Facilita la relación entre un rango y un dominio con valores de un rango específico
- Cordialidad Máxima: Permite declarar un numero amplio de relaciones en la ontología, es decir relaciones entre una clase y la otra.

La cardinalidad de la ontología oscila entre 3 y 11 (Figura 98)

[43] http://www.granddictionnaire.com

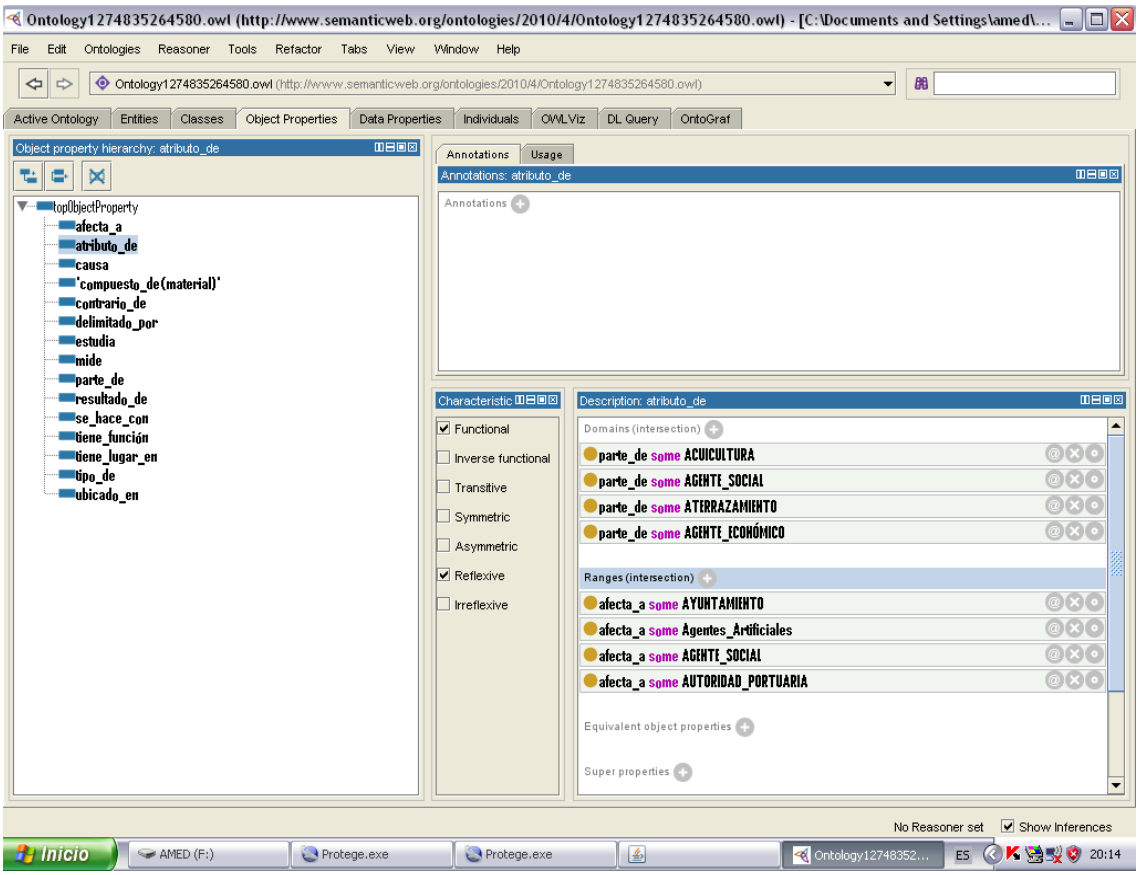



Figura.98. Relaciones de Cardinalidad

### *6.6.- Python. Lenguaje de programación*

Python es un lenguaje de script de elevada eficacia en la programación al disminuir el tecleado de códigos. Se ha convertido en una herramienta en diversas aplicaciones y sistemas. Téngase en cuenta que se usa en todos los sistemas Unix actuales (GNU/Linux, familia BSD, Solaris, etc.) y en casi todos los sistemas operativos que se usan en la actualidad. Supera en eficiencia a Bash y a Perl. Python posee un lenguaje scrip de alto nivel de interpretación y un dinamismo que resume toda la potencia de los lenguajes orientados a objetos. Python posee varios requisitos por los que se le hace elegible en este proyecto:

- Por ser un lenguaje de script solo necesita de un intérprete alojado en el sistema para su ejecución, cualidad que minimiza el proceso de llamada del programa desde el programador de tareas.

- Permite manejar RDF, pues posee bibliotecas autóctonas y puede conectarse con otras desarrolladas en otros lenguajes. Es RDFLib[44] la biblioteca de RDF más desarrollada y madura para Python, debido a su nivel elevado de consultas en SPARQL, sus potencialidades de serialización, su consistencia semántica y solidez estructural.
- Facilita la interacción del sistema con otras bibliotecas como Sparta[45] si se necesitara buscar otros conceptos que no tiene el RDF.

Se trata de un lenguaje muy expresivo y compacto, lo que conlleva directamente que se requieran menos rutinas que con otros lenguajes sin menosprecio de su capacidad, al ser de alto nivel. La sintaxis empleada en Python es fácilmente legible –en comparación a otras-, lo que facilita la lectura. Es, además, un mecanismo dual, al poder usarse como lenguaje imperativo procedimental o como lenguaje orientado a objetos, lo que le confiere una elevada flexibilidad. Entre otras ventajas, destacamos que emplea por defecto listas, diccionarios, tuplas, etc.; soporta la herencia múltiple (al contrario que Java); se puede trabajar con una gran cantidad de módulos, que amplían las opciones de trabajo; permite ejecutar procesos en paralelo, algo fundamental en este proyecto para no ralentizar el sistema; emplea plugins en C o C++; y, por último, su intérprete es gratuito.

### 6.6.1.- C#

Es un lenguaje de programación que patentiza Microsoft, que ha sido normado ECMA[46], debido a su uso intensivo en su plataforma. C Sarp posee SemWeb[47], una biblioteca, que aunque se ha desarrollado, aún es deficiente para aplicaciones complejas con RDF y SPARQL. Otras posibilidades también se ofrecen en los bindings de C#, pero son ineficaces por tener un rendimiento bajo para aplicaciones de alto nivel. Aún se necesita una biblioteca óptima para RDF desde C#. El desarrollo de herramientas a partir de estos lenguajes para la Web Semántica ha estado relegado a lo que está establecido por W3C.

[44] http://rdflib.net/
[45] http://www.mnot.net/sw/sparta
[46] http://www.ecma-international.org/publications/standards/Ecma-334.htm
[47] http://razor.occams.info/code/semweb/

### 6.6.2.- Ruby

Ruby clasifica como un lenguaje de programación de alto nivel orientado a objetos muy utilizado por la aparición del framework web Ruby on Rails hace más de 20 años. Ruby posee una sintaxis de base lingüística de alta similaridad con la Perl. Además, Ruby posee dos intérpretes muy desarrollados: Ruby y JRuby, los cuales pueden utilizarse en casi todas las plataformas al uso. Aunque ha evolucionado sus presupuestos son casi ineficientes para lo que pretende hacer en esta tesis.

### 6.6.3.- Perl

Clasifica como uno de los lenguajes script de más solidez en el mundo. Sus formas de ejecución están sustentadas en el paradigma de programación interpretativa estructurada y orientada a objetos, logrando una lógica funcional que facilita la construcción de aplicaciones de un elevado nivel a través de RDF. Su uso está extendido tanto a sistema UNIX como Window. Posee bajos requerimientos. Su biblioteca principal RDF Store[48] posee un API de calidad para interrogar RDF al igual que una para Window recomendada por W3C. Se debe aclarar que existe un culmen de bibliotecas y módulos que hacen posible el desarrollo de Web Semántica en Perl. Una de ella es MessageParser, que posee una dificultad que la hace excluible: poseer una sintaxis muy encriptada que dificulta leer y rescribir sus códigos.

### 6.6.4.- Bibliotecas y Herramientas y Lenguajes de Consulta

Son disímiles las bibliotecas que existen para el tratamiento RDF, sin embargo no todas sirven para los fines que se proponía esta investigación. A continuación mostramos las bibliotecas que se utilizaron en la investigación:

#### 6.6.4.1.- RDFLib

Es una de las bibliotecas existentes dentro de Python para manejar RDF. Constituye la biblioteca más importante pues todo el proceso de programación se basa en ella, aunque su documentación y basamento teórico no es bueno, lo que hace que aunque existan muchos desarrolladores trabajando con ella, sea

[48] http://rdfstore.sourceforge.net/

muchas veces inutilizable. Esto indica que aún tiene muchos aspectos que mejorar y muchas cosas para documentar.

#### 6.6.4.2.- PyQT4

Es un wrapper cuya sencillez facilita la construcción de interfaces visuales para los usuarios a partir de Python. En esta investigación en concreto se utilizó para diseñar la interfaz de búsqueda, la de los usuarios y la interfaz de entrega de resúmenes. Conjuntamente con Qtdesigner, facilita la construcción de visualizaciones de carácter dinámico y estático.

#### 6.6.4.3.- Protégé

Pocas son las herramientas para la edición de ontologías: Protégé[49], SweDE[50] y SWOO[51]. La popularidad de cada una está en dependencia de lo que se desee hacer en la implementación. Se ha optado en esta investigación por Protégé por las siguientes razones:

- Ninguna de las herramientas de edición de ontologías facilitan la actuación como grandes dominios conceptuales.
- Implementación amigable y fácil de utilizar por los usuarios.
- Facilidades para exportar la ontología a diversos formatos como turtle, xml-rdf, etc.
- Es gratuito y posee un razonador de alto nivel.

Se reconoce que existen otros editores de ontologías como Topbraid, etc. Muchos de ellos son de pago e inaccesibles desde Cuba.

#### 6.6.4. 4.- XML Marker

XML Marker es un Editor XML que facilita la visualización de datos en XML en forma de tabla-árbol-y-texto, detallando las jerarquías de los datos de XML en forma de tablas. Cualquier etiqueta que se cree en XML Marker puede ser organizada en columnas. Es un software con un nivel de implementación eficiente si se analiza la memoria física y virtual del ordenador, es capaz de

[49] http://protege.stanford.edu/plugins/owl/
[50] http://owl-eclipse.projects.semwebcentral.org/
[51] http://www.mindswap.org/2004/SWOOP/

navegar a través de ficheros muy extensos si crear complicaciones de trabajo. Permite la depuración de los registros provenientes de texto plano. Facilita la construcción de DTDS minimizando los errores y corrigiendo los desbalances sintácticos. Es posible usar las capacidades gráficas de XML Marker para determinar los posibles errores en la programación.

### 6.6.5.- SPARQL: lenguaje de consulta

SPARQL es un lenguaje de consulta desarrollado para consultar bases de conocimiento, no es el único lenguaje de consulta de ontologías, pero se decidió usarlo en este software por sus posibilidades sintácticas, que lo hacen el más completo de los lenguajes al uso, además es el lenguaje que está siendo normado por W3C y el (RDF Data Access Working Group). Otra de sus bondades es la similitud de su sintaxis con otros lenguajes de consulta relacionales como SQL, pues en sus álgebras[52] existe una similitud extraordinaria. Además SPARQL es un lenguaje que tiene una API para Python, que es lenguaje de programación utilizado en esta tesis.

## 6.7.- Metodología de Implementación del Sistema

Abundan metodologías para el desarrollo de los sistemas, en este caso se utilizó la metodología de diagramación UML por ser un proceso que permite la clarificación y la especificación de diversas tareas en el sistema entre las que se encuentran: especificación de casos de uso, actores, límites del sistema, requerimientos, vista de distribución, vistas de proceso y las vistas de implementación.

### 6.7.1.- Especificación de casos de uso

Este segmento de la investigación contiene la descripción de los casos de usos del sistema. Los casos de uso sirven para determinar las acciones que pueden hacer cada uno de los usuarios en los roles que se les asigne en el sistema, así como las tareas que debe hacer el sistema y los nexos de este con su entorno. La utilidad de los casos de uso se centra en las actividades de análisis, diseño y test. Es importante aclarar que no todos los requisitos funcionales del sistema

[52] http://www.w3.org/2001/sw/DataAccess/rq23/rq24-algebra.html

pueden ser obtenidos tan solo con los casos de uso, por lo que más adelante se declararán requisitos de rendimiento y fiabilidad (Capítulo 7).

#### 6.7.1.1.- Actores

Se define al actor dentro de un sistema de software a un individuo o sistema externo que realiza diversos roles interpretados bien sea por el sistema y por él mismo (Jacobson et al., 2000). A continuación se declararán los actores del sistema, así como una breve descripción de cada uno de ellos:

- **Usuario:** Es la persona que interactúa con el sistema y las base de conocimientos, no maneja ni actualiza datos, los utiliza para obtener información a través de los consultas.
- **Administrador:** Se encarga de la administración de los servicios de información y el servidor. Sus tareas principales son: programar las actualizaciones y constatar que las mismas se hayan realizado correctamente.
- **Bibliotecario:** Es capaz de revisar las necesidades de los usuarios descritas en el formulario de entrada para procesar (marcar) manualmente los documentos y las revistas que se soliciten. También es encargado de subir estos documentos a la base conocimiento.

#### 6.7.1.2.- Límites del Sistema

En la figura que se muestra a continuación aparecen los principales casos de uso del sistema y los límites que se establecen entre cada uno de los roles, más adelante se especificarán aún más los casos de uso del sistema (Ver figura 99) (Anexo 37).

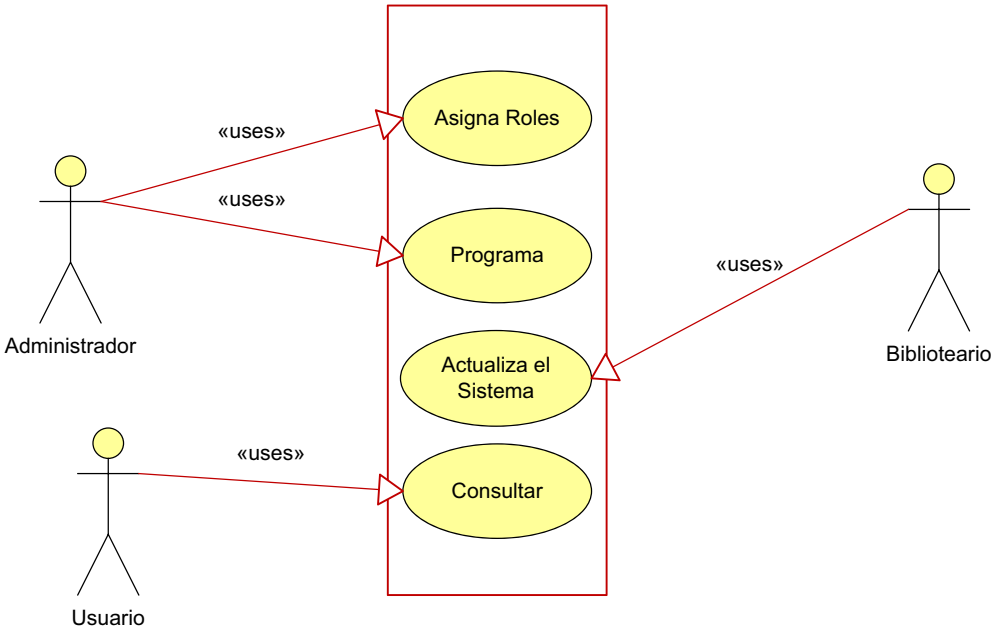



Figura 99. Límites del Sistema

Como puede verse en la figura se pueden distinguir 4 grandes segmentos de casos de uso:

- Asignar Roles
- Programar
- Actualizar los contenidos del Sistema
- Consultar

Del caso de uso anterior se especifican algunos otros que se denotan a continuación:

- Desarrollar Nuevos Módulos
- Etiquetar documentos
- Subir documentos al sistema
- Realizar Búsquedas
- Resumir un documento
- Visualizar Información
- Obtener Estadísticas

A continuación se especifican cada una de los casos de uso de forma detallada.

**Probar Confiabilidad de los Datos**

- **Descripción:** Este caso de uso representa la labor que el usuario administrador debe realizar para hacer nuevos módulos o actualizaciones del sistema.
- **Flujo de eventos:** El caso de uso comienza cuando el usuario administrador recibe una propuesta de actualización de parte del bibliotecario.
- **Precondiciones:** Es necesario disponer de un servicio de mailbox para recibir los mensajes.
- **Postcondiciones**: Ninguna detectada.

**Etiquetar documentos**

- **Descripción:** El proceso de etiquetar documentos consiste en marcar en xml la superestructura de un artículo científico así como su estructura retórica.
- **Flujo de Eventos:** El caso de uso se inicia cuando el bibliotecario recibe un perfil con una solicitud de información nueva.
- **Precondiciones:** Se necesita lograr que el sistema reconozca cuando se ha introducido un nuevo formulario y que le envíe mediante mailbox un mensaje el bibliotecario para que revise la solicitud y la procese.
- **Postcondiciones**: Ninguna detectada.

**Subir documentos al sistema**

- **Descripción:** El bibliotecario recibe el nuevo perfil de usuario, localiza los documentos en Internet, los marca en la ontología y los sube al sistema.
- **Flujo de Eventos:** El caso de uso empieza cuando el bibliotecario observa que tuene un perfil de usuarios nuevo, hace la búsqueda y con

el editor de ontología pone los documentos en el sistema y los sube a la aplicación.

- **Precondiciones:** Es vital que la comunicación del usuario con el bibliotecario mediante el mailbox.
- **Postcondiciones:** No demanda ninguna.

**Realizar Búsquedas**

- **Descripción:** Permite al usuario realizar la búsqueda en el módulo de consulta.
- **Flujo de Eventos:** El caso de uso empieza una vez que el usuario se ha autentificado en el sistema y llenado el perfil de usuario. Buscar en el DublínCore, FOAF, Consultar clases y subclases en OWL a partir de diversos criterios entre los que se encuentran (autor, título, materia, fecha, tipo de documento). El Sistema almacena el texto resumido automáticamente.
- **Precondiciones:** Es importante que el usuario llene correctamente el perfil para que se lleve a cabo la búsqueda.
- **Postcondiciones:** No demanda ninguna.

**Resumir un documento**

- **Descripción:** Mediante la opción resumir un documento es posible obtener diversos resúmenes sobre artículos especializados.
- **Flujo de Eventos:** El caso de uso comienza cuando el usuario ha terminado la búsqueda y seleccionado el documento que desea resumir.
- **Precondicones:** Es vital que los usuarios hayan buscado y seleccionado antes aquellos documentos que desean resumir.
- **Postcondiciones:** No demanda ninguna.

**Visualizar Información**

- **Descripción:** Facilita la visualización de conceptos en el tesauro visual del sistema.

- **Flujo de Eventos:** Este caso de uso comienza cuando el usuario introduce un término en la interfaz de búsqueda con el objeto de visualizar sus especificaciones.
- **Precondiciones:** No demanda.
- **Postcondiciones:** No demanda.

**Obtener Estadísticas**

- **Descripción:** Facilita la consulta de elementos estadísticos en el sistema, entre ellos están la cantidad de clases, subclases, conceptos, etc.
- **Fulo de Eventos:** Comienza con la consulta específica de los elementos estadísticos del sistema.
- **Precondiciones:** No demanda.
- **Postcondiciones:** No demanda.

### 6.7.3.- Requerimientos

El software necesita necesitar de requerimientos hardware elevados, siendo capaz de ejecutarse en un procesador de como mínimo 1.3 GHz, con un mínimo de 4 Gb de memoria RAM y un Disco duro de 370 Gigas.

Los requisitos concretos (procesador de 32 o 64 bits, sistema operativo, etc.) estarán desarrollados en función de las restricciones que existan en el escenario donde se instalado.

#### 6.7.3.1.- Requisitos de documentación

Como todo sistema desarrollado bajo estándares de trabajo será necesario que se especifique los manuales del uso del sistema, para ello se han exigido los tres documentos siguientes:

- Manual técnico en el que se recoja toda la información que se demanda para extender parcial o totalmente el software.

- Manual de despliegue en el que debe describirse todos los requisitos y las acciones necesarias para la instalación del software.
- Manual de usuario: contentivo de los pasos que necesita el usuario para usar el sistema, alejado de tecnicismos y de formas ingenieriles.

Es importante destacar que todos los documentos que se usen en el sistema deben estar formateados en pdf.

#### 6.7.3.1.1- Plan de Desarrollo del Proyecto

**Estimación de Recursos**

En esta etapa del proyecto se estiman los recursos e índole material e individual necesarios para la ejecución del mismo (Ver Tablas 53-54).

**Recursos Materiales**

| Id Unidad | Descripción | Unidad de Medición | No Unidades |
|---|---|---|---|
| HW1 | Ordenador de tipo PC | unidad | uno |
| HW2 | Ordenador de tipo PC | unidad | uno |
| SW1 | S.O. GNU/Linux | unidad | uno |
| SW2 | Intérprete de Python | unidad | uno |
| SW3 | Protegé | unidad | uno |

Tabla 53. Recursos Materiales

### Personales

| Id Unidad | Descripción | Unidad de Medición | No Unidades |
|---|---|---|---|
| HU1 | Especificación de requisitos y Estudio de viabilidad. | Días | 45 |
| HU2 | Análisis y diseño | Días | 50 |
| HU3 | Desarrollo de Software | Días | 434 |
| HU4 | Documentación | Días | 24 |
| HU5 | Dirección Técnica | Días | 23 |

Tabla. 54. Recursos Personales

### Etapas del proyecto

El desarrollo del proyecto (Ver tabla 55-56 y figura 100) ha quedado delimitado en varias etapas claramente marcadas que a continuación se declaran:

| WBS | Tarea | Inicio | Fin |
|---|---|---|---|
| 1 | Especificación de requisitos y estudio de viabilidad | 2009/03/01 | 2009/05/01 |
| 2 | Análisis y Diseño | 2009/05/02 | 2009/07/01 |
| 3 | Desarrollo Software | 2009/07/02 | 2011/03/09 |
| 3.1. | | | |
| 3.2. | Diseño de la ontología | | |
| 3.3. | | | |
| | Desarrollo de Puertotex | | |
| | Desarrollo de las herramientas complementarias | | |
| 4 | Elaboración de la documentación | 2011/02/01 | 2011/03/05 |
| 5 | Dirección | 2011/03/05 | 2011/04/05 |

Tabla 55. Planificación de las Tareas

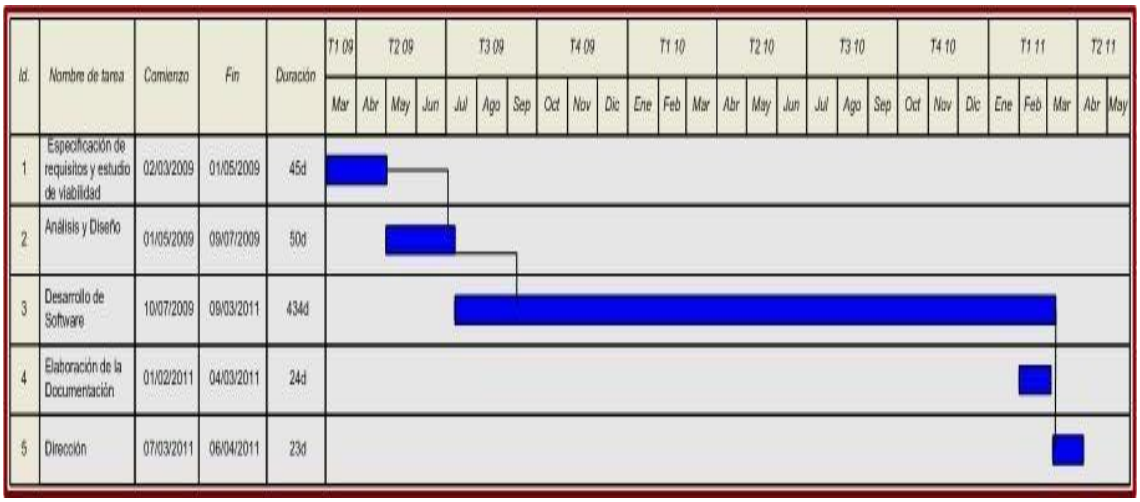



Figura. 100. Diagrama de Gantt

## Presupuesto

| ID Unidad | Descripción | Unidad de medición CUC | Precio |
|---|---|---|---|
| HW1 | Ordenador de tipo PC | CUC | 1.100 |
| HW2 | Ordenador de tipo PC | CUC | 1.100 |
| HW3 | Ordenador Portátil hp | CUC | 1.100 |
| HW4 | Ordenador tipo PC (Servidor) | CUC | 2.000 |
| SW1 | S.O. GNU/Linux | CUC | 0 |
| SW2 | Intérprete de Python | CUC | 0 |
| SW3 | Protégé | CUC | 0 |
| SW4 | XMLMarker | CUC | 0 |
| SW5 | Qt Designer 4.7 | CUC | 0 |
| SW6 | Módulo rdfLib para Python | CUC | 0 |
| SW7 | Módulo PyQt | CUC | 0 |
| SW8 | Módulo PyMySQL | CUC | 0 |
| HU1 | Especificación de requisitos y estudio de viabilidad. | CUP | 2000 |
| HU2 | Análisis y diseño | CUP | 1876 |
| HU3 | Desarrollo de Software | CUP | 14944 |
| HU4 | Elaboración de la Documentación | CUP | 1876 |
| HU5 | Dirección Técnica | CUP | 1882 |

Tabla 56. Precio

## Presupuesto final

| Descripción | Importe |
|---|---|
| Recursos hardware | 84500.00 |
| Recursos software | 0.00 |
| Recursos personales | 20578.00 |
| Total | 105078.00 |
| Beneficio Industrial 6 % | 6304.68 |
| Costos Generales 15 % | 15761.70 |
| Suma de gastos y Beneficios | 22066,38 |

Tabla 57. Presupuesto Final

### 6.7.4.-Diseño del Sistema de Resúmenes

El diseño de un sistema de software posee diversas actividades que hacen de él un una etapa única con carácter multidimensional. En la figura (Ver figura 101) se detalla la funcionalidad, organización y topología del sistema.

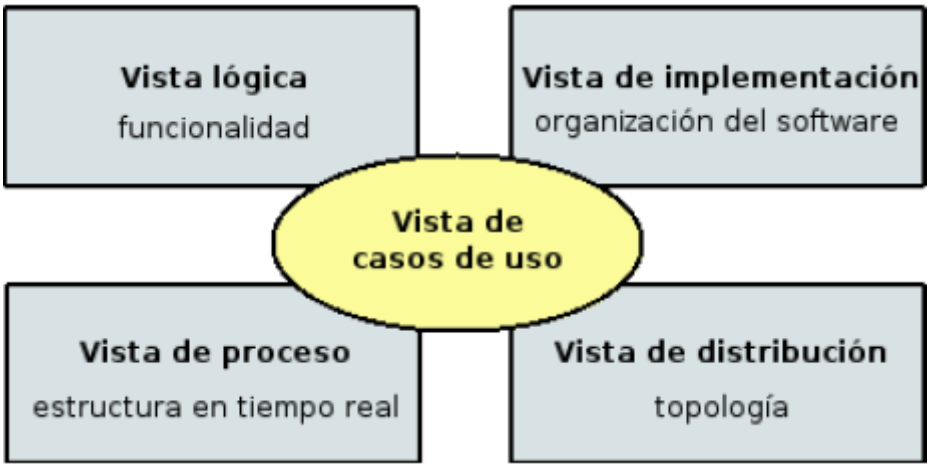



Figura 101 Procesos de Análisis de Software (Fernández López, n.d.)

- **Vista de casos de uso:** Facilita una representación gráfica amplia sobre los diversos roles que asume un usuario frente al sistema, determinando de esta forma la forma los modos de actuación del sistema y su funcionamiento.
- **Vista lógica:** Estos diagramas muestran la lógica de funcionamiento del sistema. La vista lógica se sustenta en diagramas de clases con sus respectivas relaciones capaces de mostrar diversas abstracciones que constituyen las claves en el sistema que se desarrolla.
- **Vista de proceso:** Es capaz de mostrar una graficación que maneja en tiempo real diversos componentes del sistema teniendo en cuenta

parámetros diversos entre los que se encuentran: el rendimiento, la fiabilidad, escalabilidad, integridad, organización del sistema y sincronización.

- **Vista de implementación:** Se encarga de presentar el software mediante paquetes o segmentos, para su descripción se tiene en cuenta la organización del software, la reutilización, las condiciones de desarrollo y las posibles restricciones del sistema, sobre todo las que se generan por los lenguajes de programación y las herramientas usadas en el desarrollo
- **Vista de distribución:** Permite visualizar el sistema de forma tal que este sea comprensible para los desarrolladores.

#### 6.7.4.1.- Vista de Casos de Uso

Teniendo en cuenta las especificaciones para este tema, declaradas en el segmento 6.6.1.2:

- Desarrollar Pruebas de confiabilidad con la ontología
- Etiquetar documentos
- Subir documentos al sistema
- Realizar Búsquedas
- Resumir un documento
- Visualizar Información
- Obtener Estadísticas

Se diseñarán los casos de uso en detalle a partir de este segmento de la investigación.

**Desarrollar Pruebas de Confiabilidad de Datos**

Este caso de uso especifica las acciones que debe hacer el administrador del sistema para desarrollar una modificación en el mismo (Ver figura 102).

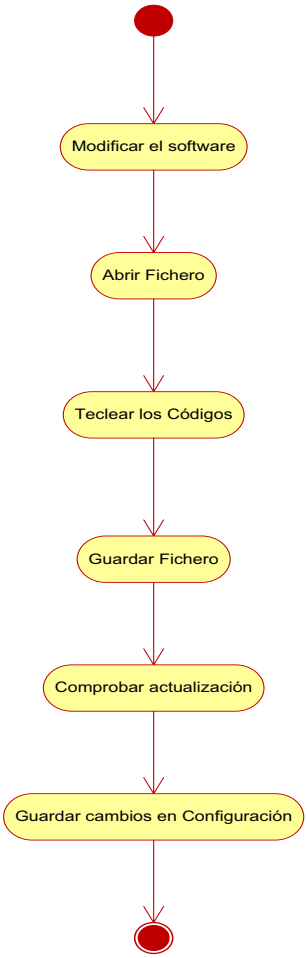



Figura 102. Diagrama de Actividad para Caso de Uso Desarrollar Nuevos Módulos

### Etiquetar documentos

Representa la acción de etiquetar nuevos documentos para que sean procesados por el sistema tanto de forma manual como de forma semiautomática y subir documentos al sistema (Ver figura 103).

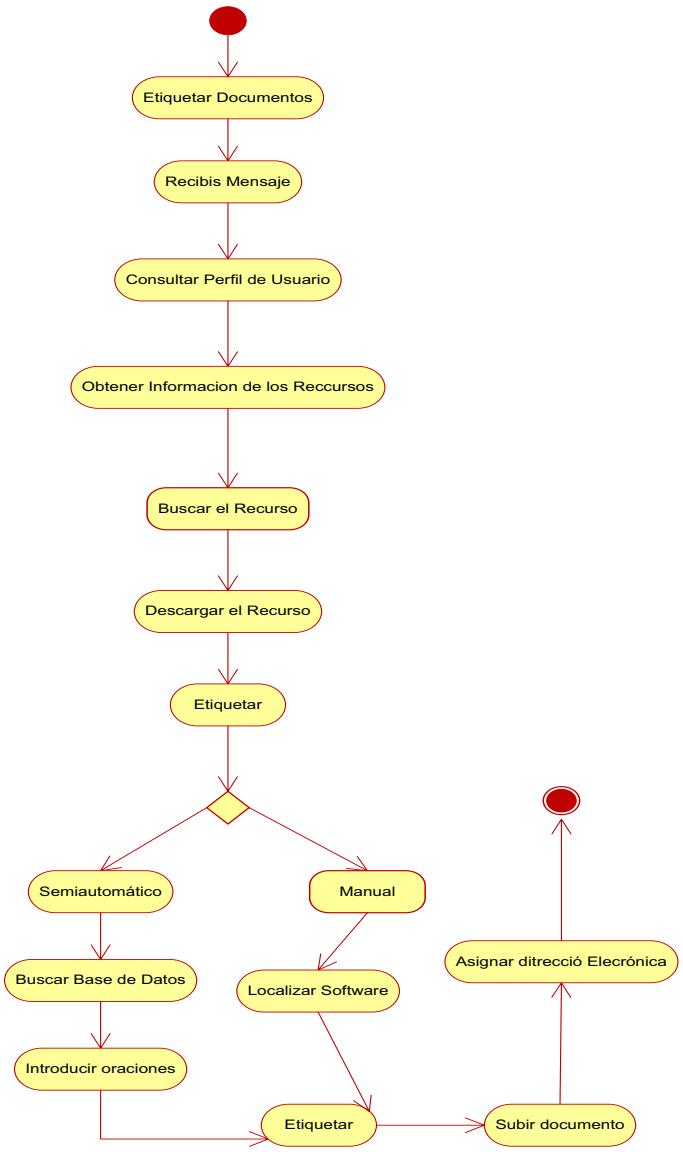



Figura 103. Diagrama de Actividad Etiquetar Documentos

### Realizar Búsquedas

Son los pasos que debe realizar un usuario determinado para buscar información dentro del sistema (Ver figura 104).

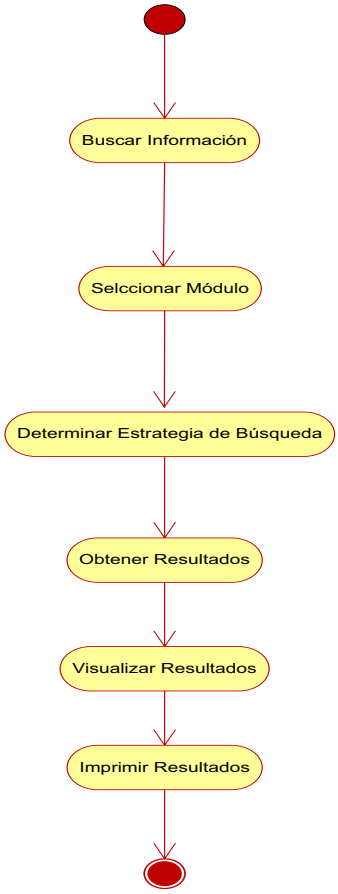



Figura 104. Diagrama de Actividad Realizar Búsquedas

**Resumir un Documento**

Acciones que realiza un usuario cando desea resumir un documento (Ver Figura 105).

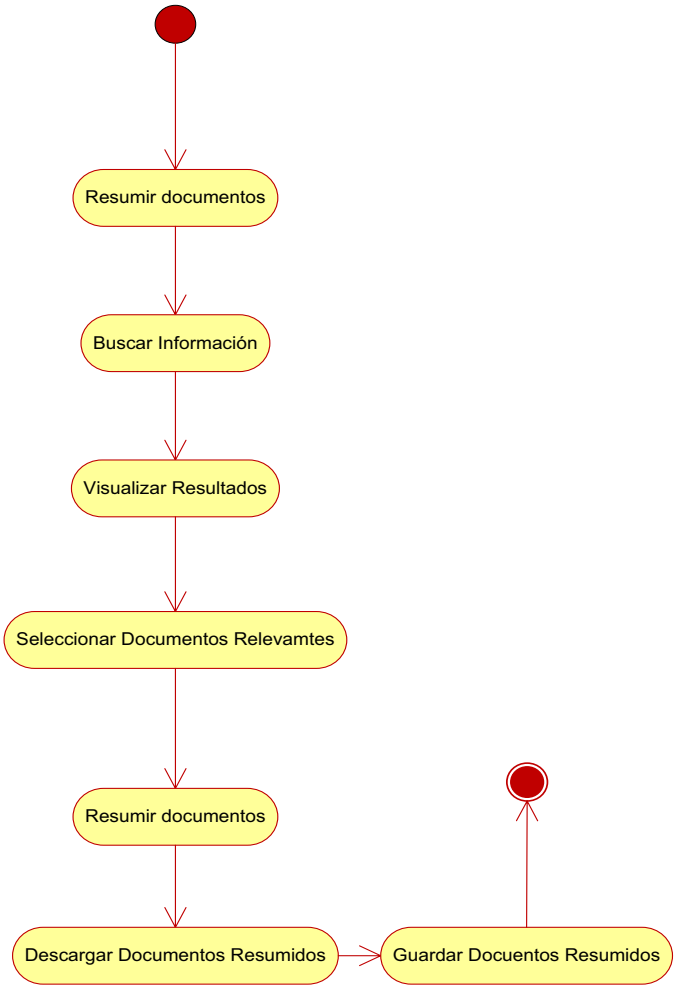



Figura 105. Diagrama de Actividad para Resumir Documentos

### Visualizar Información

Actividades que hace un usuario para visualizar conceptos en el tesauro visual (Ver Figura 106).

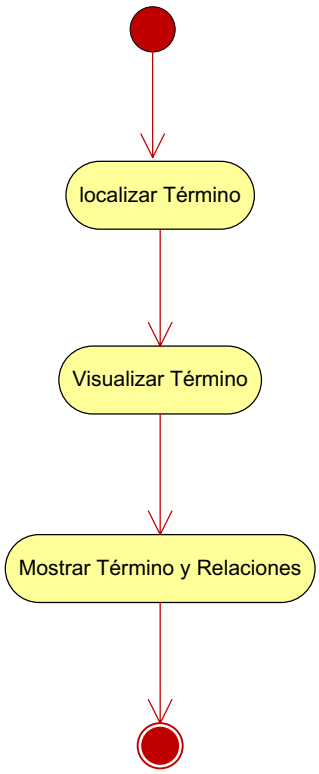



Figura 106. Diagrama de Actividad Visualizar un Término

### *6.7.5.- Vista Lógica*

El diagrama de secuencia mostrado a continuación muestra la secuencia de pasos genérica que sigue la aplicación para responder a cualquier pedido. Esta secuencia de pasos esta implementada en la estructura del controlador frontal de la aplicación y elementos asociados (Plugins, Router, Routes, Dispatcher, etc.) y asegura un mecanismo homogéneo de trabajo y una buena infraestructura de base para el tratamiento de errores y operaciones de despacho (Ver Figura 107).

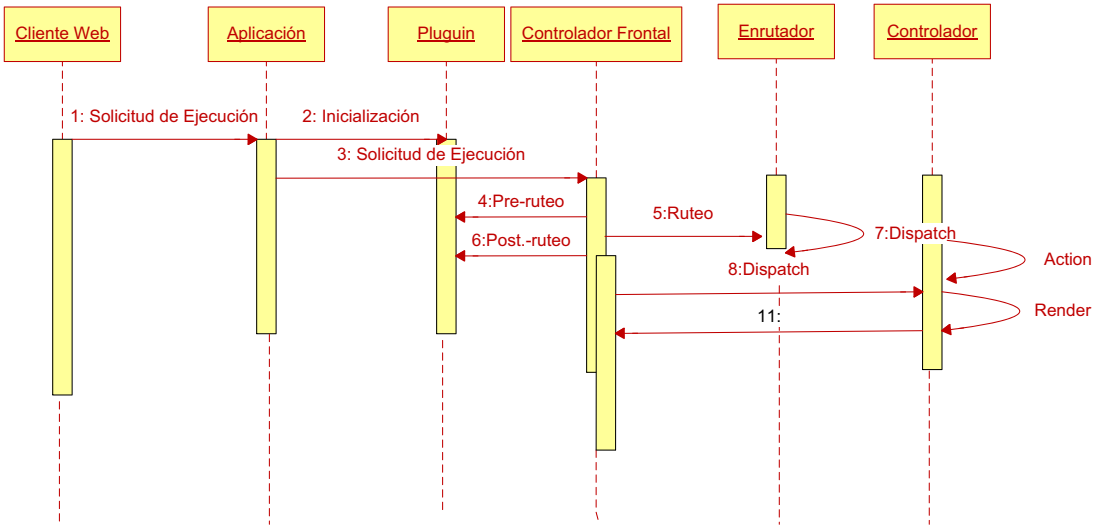



Figura 107. Diagrama de Secuencia de la Aplicación

En la figura (Ver figura) se muestra un diagrama general donde se describen el tiempo de demora de cada operación en el servidor y la secuencia en que se ejecutan las operaciones de sumarización en el Server. Es importante aclarar que existen otros procesos que no requieren de loggeado como la localización y visualización de conceptos, la búsqueda en redes sociales y la estadística. De esta forma es posible declarar como se organizan las siguientes acciones:

- **Acceder a la URL**: Localizar la dirección electrónica de la herramienta. El sistema emite una señal para que se autentifique el usuario si está previamente registrado, emitiendo el mensaje de loggeado, donde se enuncia que debe registrarse antes de autentificarse.
- **Llenar Formulario:** El usuario llena el formulario si no está registrado y el sistema emite un mensaje donde confirma que el formulario está lleno.
- **Buscar Información:** El usuario estando registrado puede buscar información, el sistema le permitirá siempre aclarar los criterios de búsqueda debido a que posee muchos mecanismos para buscar y localizar la información. Se emite el mensaje que confirma que la información solicitada registró un número X de documentos.
- **Seleccionar Información:** La Información localizada debe ser seleccionada para que se resuma o se sumarice.

- **Resumir:** Una vez seleccionada la información esta se resume y finalmente se almacena

### *6.7.6- Diagramas de Clase*

En este acápite se muestran las clases que componen el sistema y la especificación de las mismas. (Ver Figura 108)

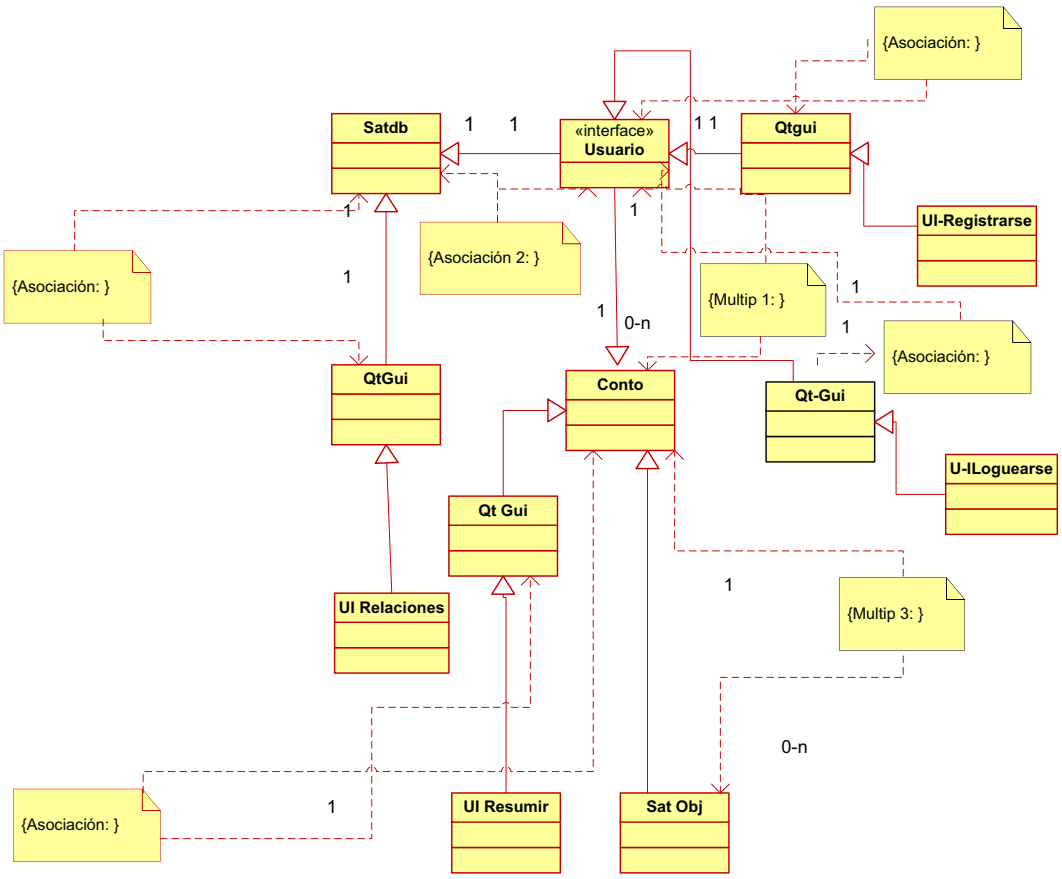



Figura 108. Diagrama de Clases (General)

#### 6.7.6.1.- Detalle de las Clases

Aquí se muestran las clases, especificando las clases, los atributos y los métodos de clases que le son asignados a cada una.

**Clase SatBD** Abstracción de Construcción de un mapa conceptual (Ver figura 109.).


Figura 109. Clase SatDB

**Clase SATObject** Abstracción de Resumidor Humano (Ver Figura 110).

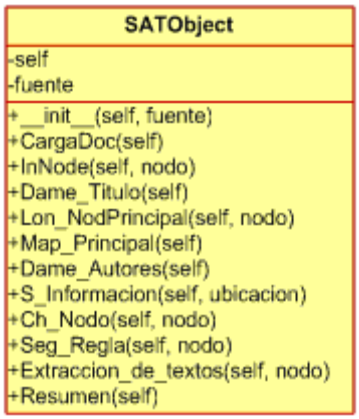



Figura 110. Clase SATObjet

**Clase COnto** Abstracción de Búsqueda y Recuperación de la Información (Ver Figura 111).

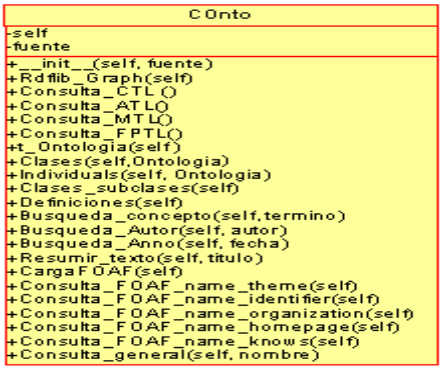



Figura 111. Clase COnto

**Clase UI_bienvenida:** Abstracción para mostrar la bienvenido al usuario al sistema (Ver Figura 112).

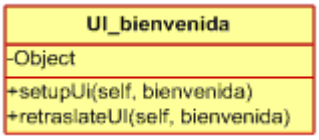



Figura 112. Clase UI_bienvenida

**Clase UI_busqueda:** Facilita la búsqueda de Información (Ver Figura 113).

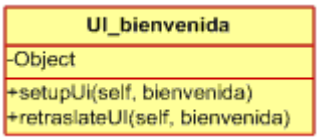



Figura 113. Clase UI_busqueda

**Clase UI_logearse:** Interviene en el logeado de los usuarios (Ver Figura 114).

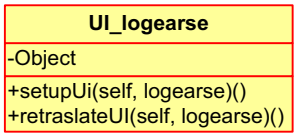



Figura 114. Clase UI_logearse

**Clase UI_usnencontrado:** Declara las acciones que suceden si un usuario no es localizado en el sistema (Ver Figura 115).

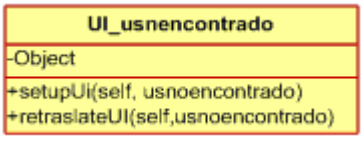



Figura 115. Clase UI_logearse

**Clase UI_principal:** Declara el menú principal (Ver Figura 116).

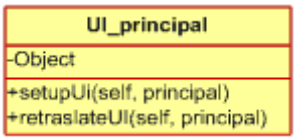



Figura 116. Clase UI_principal

**Clase UI_registrarse:** Registra a los usuarios en el sistema de acuerdo a los roles que se presentan (Ver Figura 117).


Figura 117. Clase UI_registrarse

**Clase UI_relaciones.** Especifica las relaciones entre las clases, subclases e instancias en el sistema (Ver Figura 118).

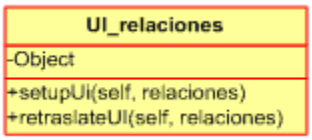



Figura 118. Clase UI_relaciones

**Clase saludo**: Declara el saludo del sistema, es una clase interfaz gráfica (Ver Figura 119).

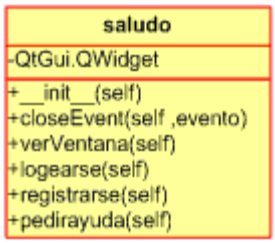



Figura 119. Clase Saludo

**Clase Búsqueda:** Interfaz gráfica de búsqueda (Ver figura 120).

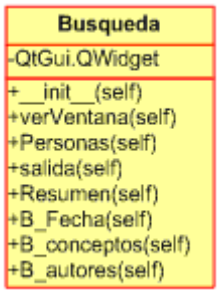



Figura 120. Clase Búsqueda

**Clase Acceder:** Interfaz gráfica para acceder al sistema (Ver Figura 121).

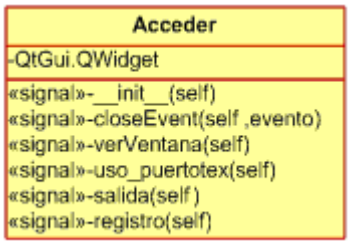



Figura 121. Clase Acceder

**Clase Usuario_no_encontrado:** Devuelve información al usuario si no es localizado (Ver Figura 122).

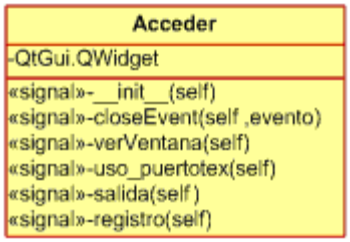



Figura. 122. Clase Usuario_no_encontrado

**Clase Menú:** Interfaz gráfica del menú principal (Ver Figura 123).

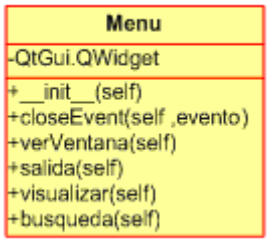



Figura.123 Clase Menú

**Clase Registro**: Interfaz gráfica de Registro de Usuario (Ver Figura 124).


Figura. 124. Clase Registro

**Clase Relaciones:** Interfaz gráfica de que muestra las relaciones (Ver Figura 125).

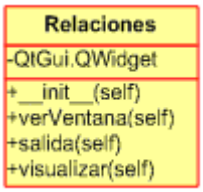



Figura. 125. Clase Relaciones

**Clase Usuario:** Fallita el registro de los datos del usuario, así como la consulta de los mismos (Ver Figura 126).

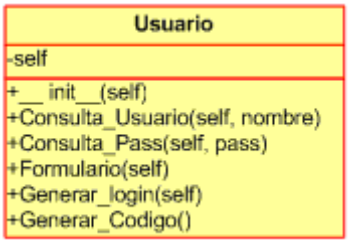



Figura. 126. Clase Usuario

Los componentes (Ver Figura 127) esenciales de este proceso se describen a continuación:

- **Ui:** Utilizado para el desarrollo de las interfaces del sistema.
- **Puertotex:** La aplicación principal la que es capaz de unir todos los procesos del sistema.
- **Rdflib:** Una librería de Python capaz de interactuar con el sistema mediante consultas en rdf.
- **PHP:** Utilizado exclusivamente para las consultas en Web.
- **MySQL:** Gestor para el tratamiento de los datos de los usuarios.
- **Interfaz:** Un medio de comunicación entre PHP y MySQL
- **Base de Datos Usuario:** Registra y almacena los datos de los usuarios que acceden al sistema.
- **Ontología:** Gran base de conocimiento, donde se encuentran los datos almacenados en formato RDF
- **TexXML:** Base de datos con textos macados en XML.
- **Minidom:** Utilizado para la lectura del documento en XML o sea es un elemento de parseo.

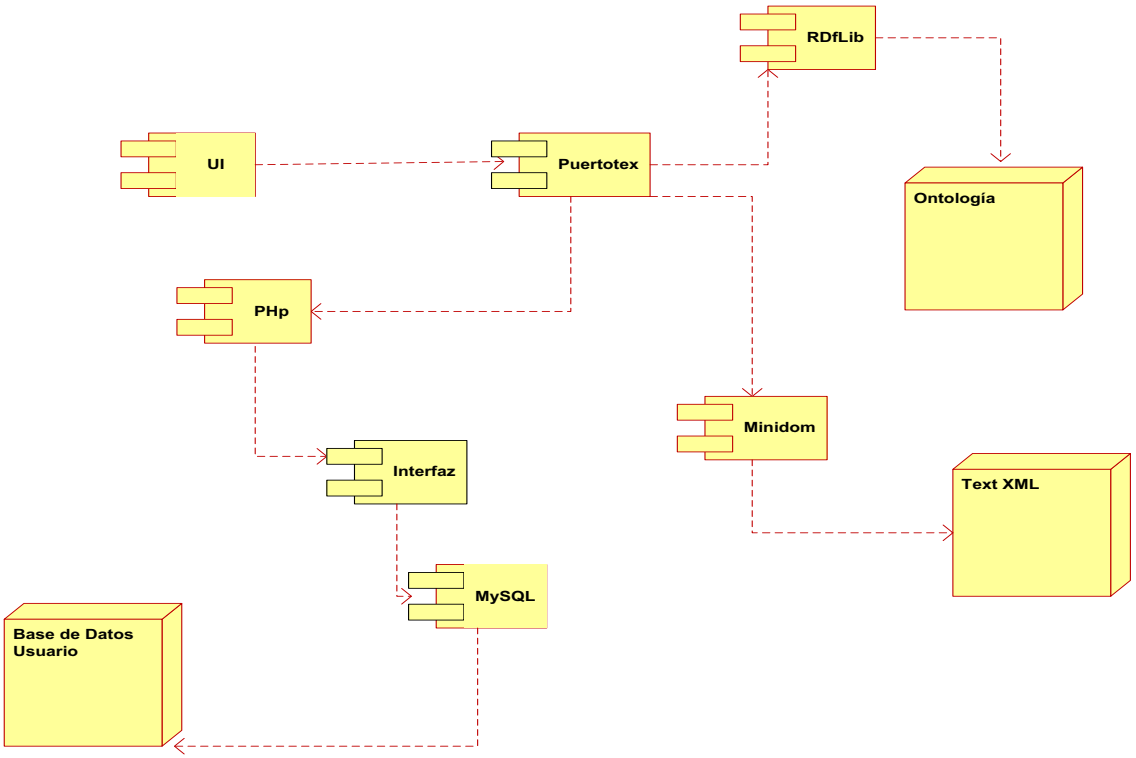



Figura. 127. Diagrama de Componentes

### *6.7.7.- Vista de Implementación*

En Python todos los elementos se describen en paquetes, por ende la vista de implementación queda establecida con satcol y los elementos que incluye es decir el paquete Satcol clases (Ver Figura 128).

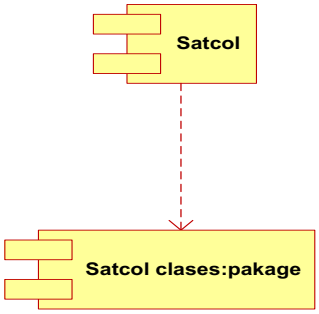



Figura. 128. Diagrama de Implementación

### *6.7.8.- Vista de Distribución*

Como puede observarse para que se desarrollo el proceso de ejecución de Puertotex se necesita de un servidos PHP, una base de datos de textos y una ontología para que la biblioteca RDFlib pueda manejar los datos y las solicitudes de los usuarios, previamente antes de ejecutar las consultas, en el servidor se demanda el trabajo con PHP, MySQL y Puertotex, en este están asociados todas las herramientas de parseo y lectura de textos (Ver Figura 129).

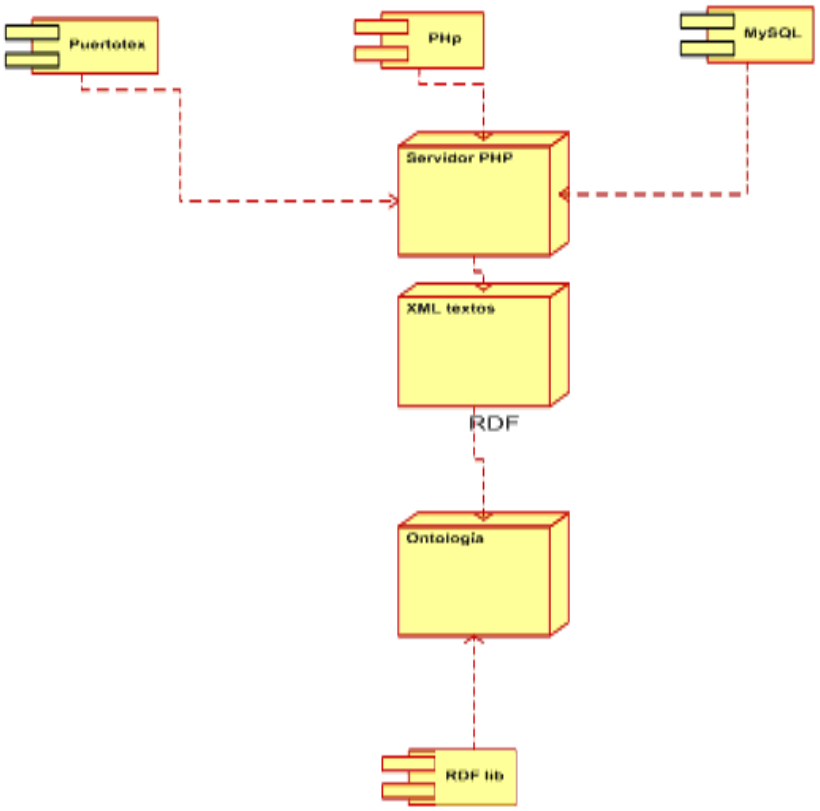



Figura. 129. Diagrama de Distribución

### *6.7.9.- Manual de Usuario*

Puerto Tex en su concepción inicial está orientado a ser un servicio Web que preste información a especialistas e investigadores de la Ingeniería de Puertos y Costas. La programación de esta herramienta es en GNU/Linux y su puesta

en operación debe limar la limitación de ser utilizada en cualquier Sistema Operativo.

Puerto Tex está orientada a tres elementos que definen su comportamiento: Visualización de términos y sus relaciones, Búsqueda y Recuperación de la Información y Resumen de textos XML previamente marcados. Se pretende dotar a especialistas de una interface agradable y sencilla en su uso que posibilite la formulación de conceptos en el dominio específico de la Ingeniería de Puertos y Costas y el resumen automático de textos basado en el Modelo TexMiner.

En el momento de conclusión de este manual, se presenta como una aplicación para realizar sobre ella pruebas que caractericen su desempeño para luego continuar en su desarrollo, esto está en concordancia con toda filosofía bajo los patrones de Software Libre.

**Requisitos Técnicos**

**Interprete de Python**

Ambiente integrado para interactuar con los Script escritos en el lenguaje de programación Python. Proporciona la interactividad con los diferentes módulos que se presentan en Puerto Tex. La instalación de este programa se puede encontrar en el servidor de repositorios para Ubuntu de la UCLV *http://10.12.1.103/ubuntu*.

Bibliotecas necesarias

- RDFlib (3.0)
- PyQt4
- qt4-Designer
- DOM
- MySQLdb

**Instalación de los Módulos.**

*RDFLib (3.0)*

La instalación de *RDFLib (3.0)* se puede descargar en http://rdflib.net/ . Es preciso tener en cuenta que el Python inferior al 3.0 aparece por defecto instalado *RDFLib* que no es la que se utiliza en la aplicación Puerto Tex, es necesario por tanto desinstalar la versión existente. Los siguientes pasos fueron los realizados para la instalación de la nueva versión de *RDFLib*.

1- Se abre una sesión del Terminal.
2- Se logea como usuario administrador del sistema.
3- Se cambia al directorio donde se encuentra descompactado la instilación de RDFLib.
4- Se ejecuta *python setup.py install*.
5- Luego de la instalación en la carpeta de *RDFLib* ubicada en (poner) se abre el fichero graph.py y en la función donde se define la consulta se agregan las variables *initNs={}, initBindings={}* y en el *return result(processor.query(query_object,initBindings,initNs))*

***Pyqt4***

Esta biblioteca de trabajo de Python 2.5 con qt4 se puede descargar desde el servidor de repositorio de la UCLV http://10.12.1.103/ubuntu. Para la instalación de la misma se teclea logeado como administrador en el Terminal:

*apt-get install pyqt4*

Otra forma de instalar esta biblioteca es mediante el Gestor de Paquete Synaptic.

*qt4-Designer*

La instalación de qt4-Designer se puede descargar desde http://riverbankcomputing.com

Luego de descargar el archivo compactado se procede de manera similar a la instalación de RDFLib (hasta el paso 4).

Otra forma de instalación es mediante descarga desde el servidor de repositorios de Ubuntu de la UCLV. Para este caso se presentan las dos vías tradicionales: Gestor de paquetes Synaptic o logeado como administrador tecleando:

*apt-get install qt4-designer*

La documentación de este programa se puede instalar tecleando:

*apt-get install qt4-doc*

*DOM*

Este módulo permite la consulta a documentos XML. La instalación del módulo viene por defecto incluida en la instalación del Python 2.5.

MySQLdb

Este es el módulo para la consulta de base de datos aplicando la consulta en lenguaje de consulta SQL. Su utilización en Puerto Tex se centra en la base de datos de usuarios que se propone en el sistema.

La instalación del módulo se realiza mediante el siguiente código tecleado desde el Terminal.

*sudo apt-get install python-mysqldb*

**Desarrollo posterior de PuertoTex**

Los usuarios que administren y desarrollen Puerto Tex deben tener los conocimientos mínimos que se presentan:

- Conocimientos avanzados de Programación Orientado a Objeto (POO).
- Conocimientos medios de Python.
- Conocer cómo funciona la biblioteca RDFLib (3.0).
- Conocer cómo funcionan las bibliotecas de trabajo con XML

específicamente DOM.

- Conocer cómo funciona la biblioteca PyQt4.
- Conocer cómo funciona la biblioteca de trabajo con MySQL.

**Obtención del Código Fuente.**

En la versión que se presenta no se pretende distribuir debido a que el software pasará un conjunto de pruebas. El código fuente de PuertoTex cuando se generalice su utilización se alojará en un Servidor destinado a prestar este servicio en las entidades científicas-docentes que lo utilicen.

**Organización del Código Fuente**

El código fuente de Puerto Tex se estructura en un directorio raíz denominado PuertoTex el cual presenta un grupo de subdirectorios que contienen los componentes que conforman el sistema (Ver Figura 130).

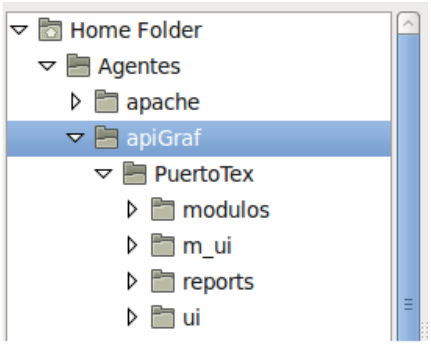



Figura. 130. Organización del Código Fuente

**Subdirectorio Módulos**

El subdirectorio Módulos contiene los tres script que definen el comportamiento de la aplicación: AgBD.py, Agente.py y Conto.py y Usuarios.py.

*AgBD.py*

Módulo que se encarga de la búsqueda de los dominios, las relaciones y los rangos de un término entrado por el usuario. Este módulo juega un papel primordial en la visualización de términos en Puerto Tex.

*Agente.py*

Módulo que realiza el resumen de un documento XML marcado.

*COnto.py*

Módulo que se encarga de la búsqueda y recuperación de la información. Se destacan en este módulo la consulta a las tecnologías RDF y OWL aplicando SPARQL.

*Usuario.py*

Módulo que se encarga del trabajo con los datos de Usuario.

**Subdirectorio ui**

El subdirectorio ui contiene todos los scrip de interfase gráfica generados por QtDesigner para la aplicación.

**Subdirectorio m_ui**

Este subdirectorio presenta los scrip que manejan la interface gráfica de Puerto Tex, permiten la interconexión de la interface gráfica con sus manejadores y con los módulos específicos de cada parte de la aplicación.

**Subdiretorio report**

En este subdirectorio se encuentran las versiones más recientes de las bases de conocimiento que conforman el sistema.

### *6.7.9.1.-Ventana Principal*

El módulo principal carga la herramienta con las opciones para logearse, registrarse en sistema y la ayuda para que llegue a completar todas las acciones que se realizan desde la ventana principal (Ver Figura 131).


Fig. 131 Ventana principal de Puertotex

### *6.7.9.2.- Registrarse*

La opción registrarse posee un diseño más amigable y declara la imagen corporativa del producto. Desde esta ventana el usuario llena una serie de datos que serán necesarios para que el sistema pueda localizar la información en la Web y generar resúmenes de los documentos localizados (Ver Figura 132).


Figura. 132. Ventana Registrarse

### *6.7.9.3.- Buscar y Recuperar Información*

La búsqueda de información en Puertotex se realiza a partir de varios criterios, entre los que se encuentran: materia, autor, año, personas. Cada fase de búsqueda conlleva diversos resultados, la fase de materia facilita la búsqueda de conceptos en la ontología utilizando los operadores booleanos y reglas de inferencia, la búsqueda por autor devuelve el titulo o la lista de títulos que se asocian a un autor y permite el acceso al texto original si se desea resumir ese título se da clic en la opción resumir y se obtiene el resumen (Ver Figura 133).


Figura 133. Búsqueda de un título para Resumirlo

### *6.7.9.4.- Resumir una Fuente*

La acción de resumir una fuente se realiza cuando el usuario ha seleccionado el titulo de la fuente, entonces el agente de búsqueda localiza el documento en la ontología, mostrando el resumen del texto, la superestructura del resumen y una dirección para que el usuario lea o descargue el documento original (Ver figura 134)

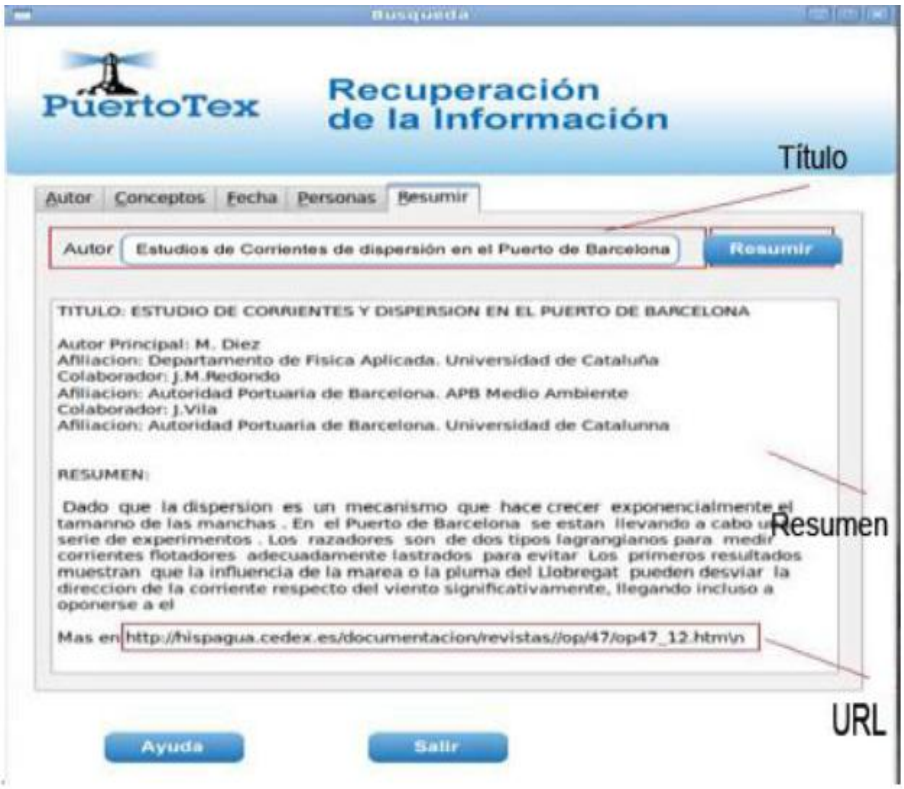



Figura.134. Opción Resumir

6.7.9.5.- Visualizar Conceptos

Una de las grandes necesidades de los sistemas de información sobre resumen basados en Web Semántica es su capacidad para generar visualizaciones de los conceptos. En Puerotex el usuario puede visualizar los conceptos y sus dimensiones semánticas como una forma der ayuda para recuperar información. El visualizador es manejable de forma dinámica y permite mover los conceptos a partir de nuevas búsquedas al igual que en Puertoterm (Ver Figura 135).

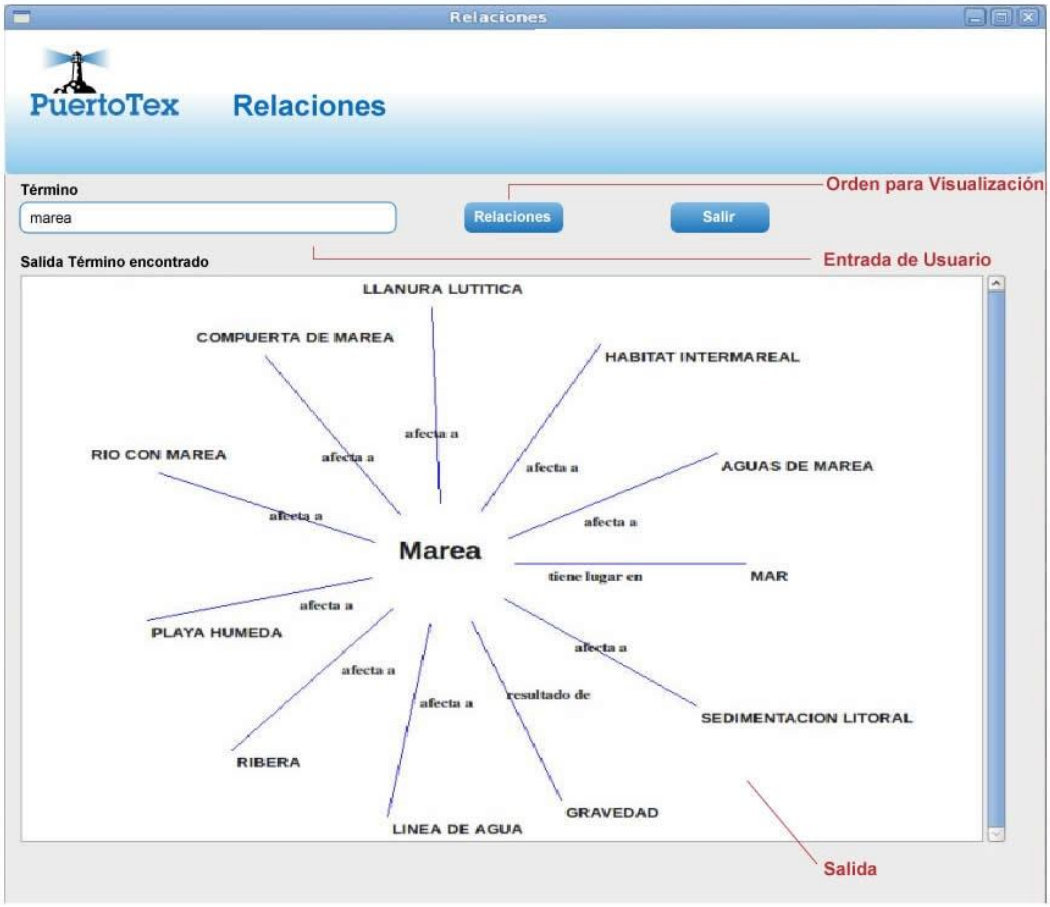



Figura. 135. Visualización de Conceptos

### *6.7.9.6.- Redes Sociales*

Puertoterm además incluye una amplia red social donde se ubican los expertos en el tema que es objeto del software. Para ello se ha implementado el estándar de FOAF para que se localice una persona, o institución, los temas que investigan, las ´personas que conoce y sus direcciones laborales, personales y su mail. El sistema de FOAF sigue la misma posición que el faceboock, por tanto realiza inferencias y mezcla todos los elementos que poseen cierta relevancia para un usuario x (Ver figura 136).

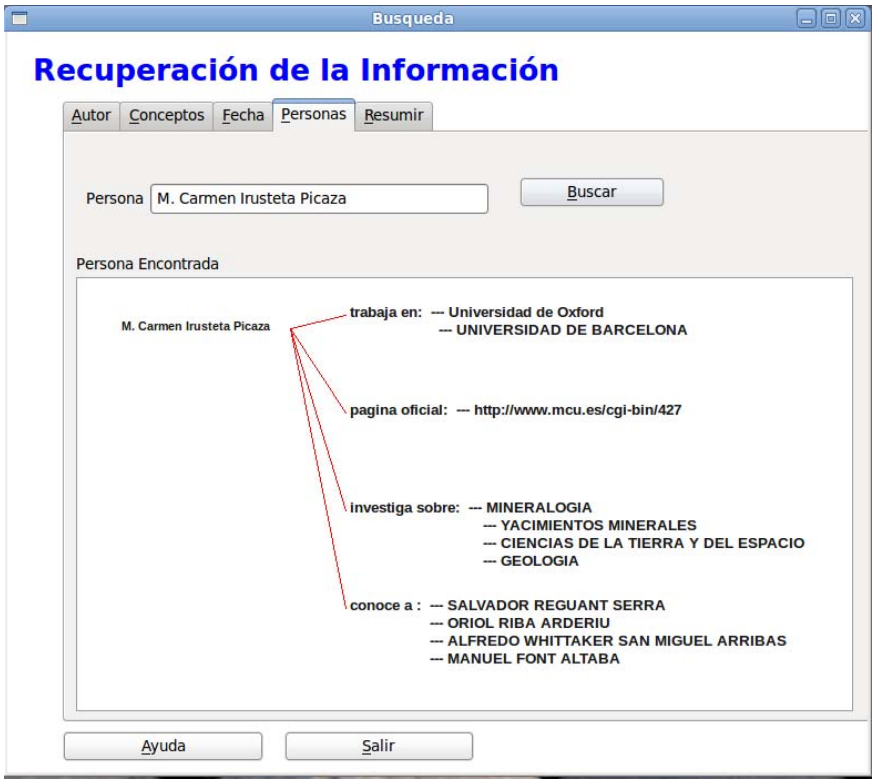



Figura. 136. FOAF

### *6.8.- Características del Servicio*

El servicio de información que resulta de la aplicación de la herramienta se denomina TEXMINER a continuación se exponen un grupo de principios tenidos en cuenta para el desarrollo del servicio, para el mismo se han tenido en cuenta algunos parámetros que permiten el desarrollo e implementación de estas actividades declaradas en Esteban (Esteban, 2006). Las necesidades de información relativas al servicio han sido declaradas en el Anexo 7 (Ver Anexo 7).

#### 6.8.1.- Planificación

Según (Esteban, 2006) la planificación de un servicio de información incluye diversos aspectos que servirán para el desarrollo del servicio, ellos son: Misión, Metas, Objetivos Específicos, Objetivos Generales, Usuarios, Prioridades, Recursos Humanos. A continuación se detallan cada uno de estos aspectos:

**Misión:** Facilitar la consulta de Fuentes de Información Especializadas en el tema de Ingeniería de Puertos y Costas mediante el resumen de artículos científicos en español e inglés.

**Objetivo General:** Resumir fuentes de información especializadas en Ingeniería de Puertos y Costas, facilitando el acceso y la selección de dichas fuentes.

**Objetivos Específicos:**

1. Seleccionar todas las fuentes de información inherentes al tema referidas por el Proyecto Puertoterm de la Universidad de Granada, España.
2. Procesar las Fuentes de Información inherentes al tema mediante técnicas de marcado.
3. Construcción de bases de Datos y Bases de Conocimiento para almacenar toda la información sobre el tema.
4. Ofrecer el servicio mediante búsqueda y selección de fuentes de información.
5. Ofrecer información con valor agregado sobre autores, instituciones.

**Usuarios**:

Todos los profesores, docentes e investigadores que se asocian al tema Ingeniería de Puertos y Costas en la Universidad Central de las Villas, Cuba. Sus necesidades de información están recogidas en el Anexo 39.

**Prioridades:**

Tendrán prioridad aquellos usuarios que sean investigadores, ellos recibirán los resultados directos de la actividad.

**Recursos Humanos:**

Los Recurso humanos que demanda este servicio son informáticos (encargados del soporte tecnológico de la actividad) y Bibliotecarios (desarrollan todo el proceso de marcado de los textos).

### 6.6.8.2.- Estrategia de Acciones para el desarrollo del Servicio

La estrategia para el desarrollo del servicio constituyó un elemento imprescindible para su desarrollo a continuación se muestra la tabla con las acciones descritas para la puesta en marcha del servicio (Tabla 58):

| Rec Humano | Acción | Responsable | Fecha |
|---|---|---|---|
| Bibliotecarios | Entrenamiento en RDF | Departamento de Control Automático | 26-2-2010 |
| Bibliotecarios | Lematización y Marcación | Departamento de Lingüística | 23-4-2010 |
| Bibliotecarios | Trabajo con Ontología | Departamento de BCI | 25-5-2011 |
| Usuario | Uso de TEXMINER | Grupo de ALFIN, CDICT | 2010-2011 |

Tabla 58. Estrategia de acciones para la implantación del servicio

### 6.8.2.1.- Divulgación

Para divulgar este servicio de información se desarrollarán un grupo de actividades m que permitan su uso en la Universidad. Seguidamente se enuncian las acciones que se realizan para la divulgación:

- Colocar en la Intranet de la Universidad el Link a los Servicios.
- Mostrar en los Departamentos de Investigación y en los Docentes las bondades del servicio.
- Ofrecer cursos y entrenamientos para el desarrollo de habilidades sobre el servicio.

## 6.8.3.- Evaluación

Para la evaluación del Servicio de Información se desarrollarán diversas actividades evaluativas que tengan en cuenta los estándares establecidos por

las Ciencias Sociales y la Ciencias de la Computación. Se exponen a continuación aquellos parámetros que servirán a la evaluación del servicio:

- Validación Por Rouge: Uso de Rouge para determinar la similitud en n-gramas de un texto fuente contra una colección de entrenamiento.
- Evaluación de los textos utilizando el criterio de lingüistas.
- Valoración de la Usabilidad: Estudio de las potencialidades de la aplicación en la práctica.


# Referencias Bibliográficas

***Referencias Bibliográficas***

ARROYO, M., AGUIRRE, J., BAVERA, F. & NORDIO, N. 2008. Jtlex Un Generador de Analizadores Léxicos Traductores. Argentina.

BELTRÁN, R. A. 2007. *Adaptación de contenido para la Web Semántica: Anotaciones semánticas, estado del arte* [Online]. Bogotá. Available: www.unal.col [Accessed 16.abril 2008].

CORRALES DEL CASTILLO, J. M. 2008. *Modelo de Servicio Semántico-Difuso de Difusión Selectiva de la Información para bibliotecas digitales.* PhD., Universidad de Granada, España.

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DINGLI, A. 2003. Melita : manual Vr 1.0. Universidad de Shefeld, Department of Computer Science.

DOMÍNGUEZ, S. 2011a. Calculuscopora. beta ed. Santa Clara, Universidad Central"Marta Abreu" de las Villas: Departamento de Ingeniería Automática.

DOMÍNGUEZ, S. 2011b. FOXCORP. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de Ingeniería Automática.

DONÉS ROJAS, R. & ORTIZ RODRÍGUEZ, C. 2006. El Proceso de Anotación Semántica en Framenet en Español. *In:* LLAMAZARES, M. V. (ed.) *Actas del XXXV Simposio de la Sociedad Española de Lingüística*

ESTEBAN, M. A. 2006. Planificación, diseño y desarrollo de servicios de información digital. *In:* TRAMULLAS, J. & GARRIDO, P. (eds.) *Software libre para servicios de información digital.* Madrid: Pearson Prentice Hall.

FABER, P. & MAIRAL-USÓN, R. 1999. *Constructing a lexicon of English verbs,* Berlín, Mouton de Gruyter.

FERNÁNDEZ BREIS, J. T. 2003. *Un Entorno de Integración de Ontologías para el Desarrollo de Sistemas de Gestión de Conocimiento* Tesis Doctoral, Universidad de Murcia.

FERNÁNDEZ LÓPEZ, S. n.d. *SWAML, publicación de listas de correo en web semántica* Tesis de fin de Carrera, Universidad de Oviedo.

FONER, L. N. 1996. What's an Agent, Anyway? Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. *In:* FRANKLIN, S. & GRAESSER, A. (eds.) *Third International Workshop on Agent Theories, Architectures and Languages.* Springer-Verlag.

FRANKLIN, S. & GRAESSER, A. 1996. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. *In Third International Workshop on Agent Therories, Architectures and Languages:.* Springer-Verlag.

HANDSCHUH, S. & STAAB, S. 2002. Authoring and Annotation of Web Pages in CREAM. *WWW2002.* Honolulu, Hawaii, USA: ACM.

HAYES-ROTH, B., PFLEGER, K., LALANDA, P., MORIGNOT, P. & BALABANOVIC, M. 1995. A Domain-Specific Software Architecture for Adaptive Intelligent Systems. *IEEE Trans. Sofware Eng* 21**,** 288-301.

HUFFMAN, D. A. 1952. A Method for the Construction of Minimun-Redundancy Codes. *Proceedings of the I.R.E.*

JACOBSON, I., BOOCH, G. & RUMBAUGH, J. 2000. *El Proceso unificado de software,* Madrid, Addyson Wesley.

KROEKER, L. 2004. Seeing data: new methods for understanding information. *IEEE computer graphics and applications,* 24**,** 6-12.

LEIVA, A. 2008. *Metodología para la extracción y desambiguación de textos científicos.* Tesis de Maestría, Universidad de la Habana.

LEIVA, A., SENSO, J., DOMÍNGUEZ, S. & HÍPOLA, P. 2009. An Automat for the semantic processing of structured information. *In ISDA 9na International Conference of Desing of Software and Aplicactión.* Italia, Pissa: IEEE.

LÓPEZ-HUERTAS, M. 2008. Organización y representación del conocimiento: curso de doctorado. La Habana: Universidad de la Habana.

MAES, P. 1994. Social Interface Agents: Acquiring Competence by Learning from Users and Other Agents. *Working Notes of the AAAI Spring Syposium on Software Agents*, 71-78.

MOREIRO, J. 1996. La Técnica del resumen científico. *In:* LÓPEZ YEPEZ, J. (ed.) *Manual de Información y documentación.* Madrid: Ediciones Pirámide.

MOREIRO, J. 2004. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural,* Madrid, Ediciones Trea.

MOREIRO, J. 2006. *El resumen científico en el contexto de la teoría de la documentación. Texto y descripción sustancial* [Online]. Madrid. Available: Disponible en:http://www.ucm.es [Accessed 26.octubre 2006].

NOVAK, J. 1991. Concept maps and vee diagrams: two metacognitive tools to facilitate meaningful learning. *Instructional Science,* 19, 1-25.

OECD 2000. *Knowledge management in the leanrnig society,* Paris, OECD.

RUSSELL, S. J. & NORVIG, P. 2003. *Artificial Intelligence : A Modern Approach.,* Upper Saddle River, N.J., Prentice Hall/Pearson Education.

SENSO, J. 2008. *Descripción e intercambio en la web semántica* [Online]. Granada: Universidad de Granada. Available: http://documentacion.ugr.es [Accessed 28. septiembre 2010].

SENSO, J. 2009. *Representación del conocimiento en la Ingeniería de Puertos y Costas.* Proyecto Investigador, Universidad de Granada.

SENSO, J. A., MAGAÑA, P. J., FABER-BENITEZ, P. & VILA, M. M. 2007. Metodología para la estructuración del conocimiento de una disciplina: el caso de PuertoTerm. *El profesional de la Información,* 16, 591-604.

SMITH, D. C., CYPHER, A. & SPOHRER, J. 1994. Programming Agents without a Programming Language. *Communications of the ACM,* 37, 55-67.

VIRDHAGRISWARAN, S. 1994. Heterogeneous Information Systems Integratin - an Agent Messaging Based Approach. *Third International Conference*

*on Information and Knowledge Management (CIKM'94).* Gaithersburg, Maryland: National Institute of Standards and Technology.


# Capítulo VII

## Análisis de los resultados de la implementación del modelo

# Capítulo 7: Evaluación de los Resultados de la Aplicación del Modelo

## 7.1.- Introducción

Los modelos de Minería de Texto deben ser evaluados en su totalidad. Generalmente en la bibliografía se observa como la evaluación en estos modelos va dirigida exclusivamente al texto resultante y a la recuperación de la información. Los grandes problemas en la evaluación de la calidad de los modelos sumistas están centrados en tres puntos: Escasa participación de los usuarios en su evaluación, falsa Evaluación de sus componentes, Monodimensionalización de la evaluación. Estas tres posturas han conseguido evaluaciones parciales de la calidad de los modelos y por ende un análisis erróneo de los mismos. En esta tesis se parte del precepto de que el modelo se evalúa desde su concepción léxica hasta su posición pragmática (puesta en marcha). El autor teniendo en cuenta las características de TEXMINER (Ver Capítulo 4) ha generado dos métodos de evaluación una para Corpus y otro para la Ontología, a su vez ha utilizado Rouge para evaluar los resúmenes y ha dado un espacio a la investigación cualitativa al ofrecer criterios de evaluación del sistema a los usuarios que interactúan con él (Anexo 9 y 10). Este capítulo se divide en 5 acápites esenciales: Evaluación de Corpus, Evaluación de Ontologías, Evaluación de Resúmenes, Evaluación de la Herramienta y Criterio de Expertos.

## 7.2. Evaluación de Corpus

La evaluación de un texto para su utilización en sistemas de Representación y Organización del Conocimiento ha sido uno de los elementos más controversiales en las investigaciones en el terreno de la Minería de Texto, la Recuperación de la Información y la Ciencia de la Información. Estas disciplinas junto a la lingüística se encargan de la confección de corpus terminológicos que sirven a la construcción de diccionarios terminológicos,

ontologías, topic-maps, etc. Muchos autores enfatizan que el desarrollo de sistemas de Información como: Ontologías, Resumidores Automáticos, Sistemas de Clasificación de Textos, etc., tienen que estar precedidos por la evaluación de Textos. Sin embargo, estos criterios son muchas veces desoídos y desconocidos por los investigadores que actúan en este terreno.

La calidad de muchos software desarrollados para la traducción, la desambiguación y la extracción automática de textos ha disminuido debido a la ausencia de posturas de evaluación de textos o corpus. Por una parte, la evaluación de textos apunta hacia una paradigma multidisciplinar donde se insertan la psicología cognitiva, la lingüística, y lexicometría**.** Por otra han aparecido modelos de evaluación generados para tipos específicos de textos, que han de ser insertados en determinados software o herramientas. Dichos modelos de evaluación han estado orientados a varios aspectos como: comprobación de algoritmos, validación de observaciones, construcción de herramientas acústicas, detección de similaridad caligráfica, etc. Sin embargo, los modelos de evaluación de textos con vistas a construir sistemas de Minería de textual y Ontologías de dominio no abundan en la literatura y los que existen se han encargado de la parte matemática del asunto, dejando detrás el componente cognitivo, usado en las ontologías y las herramienta de minería de texto. Teniendo en cuenta esta situación en esta tesis se propone una metodología para evaluar colecciones de textos que sirvieron de base a la construcción de la Ontología Ontosatcol y del Sistema de resumen PUERTOTEX, utilizando para ello el corpus del proyecto Puertoterm

### 7.2.1. – Metodologías de Evaluación de Corpus y Algoritmos de Evaluación de Corpus

Las técnicas de evaluación de Corpus pueden dividirse en tres vertientes: la Computacional, la Lingüística y sociológica y la de la Ciencia de la información. La presencia de todos estos campos de estudio en la evaluación de corpus, hace muy compleja la clasificación de los mismos. El autor ha agrupado las

técnicas teniendo en cuenta que la mayoría de las prácticas internacionales se mueven en el terreno de los métodos que aquí se exponen:

#### 7.2.1.1. - Clúster

La variedad de formas de agrupar texto ha sido tal que es muy complejo para el autor encontrar una clasificación que pueda describir todos los procedimientos metodológicos que existen para el desarrollo de grupos. Debido a que esta tesis no se dirige hacia este objetivo el autor solo se detendrá en la explicitación de determinados algoritmos de agrupamiento y se explicarán aquellos métodos cuya aplicación es frecuente en investigaciones sobre el tema, además ya se declaró en el Capítulo 2 un acápite dedicado a este proceder.

Se clasifican como algoritmos jerárquicos, a aquellos que forman particiones y se centran en patrones pequeños o prototipos simulados. Estos se inician con una división o participación de entrada que puede ser aleatoria o no y culminan con un refinamiento.

Los algoritmos jerárquicos construyen una estructura jerárquica de los objetos. En el mismo terreno de los algoritmos jerárquicos aparecen los aglomerativos, los cuales consideran que tanto objetos como grupos son la misma cosa. Esto hace que objetos y grupos sean considerados como elementos similares. Por su parte los algoritmos divisivos (top-down) parten del supuesto de la concurrencia de un grupo general, por tanto todos los objetos se encuentran incluidos en él. Este algoritmo segmenta los grupos hasta tener en ellos un solo objeto. Si se evalúa la calidad de los métodos jerárquicos aglomerativos se pueden apreciar dificultades para examinar los objetos desde la teoría de grafos. A esta clasificación se suman como ejemplos Single-Linkaje y Group Average. La característica esencial de estos métodos es la estructuración de los grupos en una estructura taxonómica, donde cada grupo contiene otro. Es muy riesgoso el uso de estos métodos que elaboran particiones, pues como dice (Arco, 2008) se crean particiones en que usualmente se requiere especificar el número de grupos como un parámetro de entrada.

Otro tipo de clúster es el basado en densidad (density-based clustering) el cual agrupa objetos de determinada vecindad a un conjunto de datos utilizando las propiedades de la densidad. A este tipo de algoritmo pertenecen DBSCAN y MajorClusf, Recurrencia de Lance & Williams y Expectation-Maximization (EM)

Otras clasificaciones declaran algoritmos iterativos a aquellos que describen itinerarios o distancias. Estos se dividen en: Ejemplar Base y Commutación Base. A este tipo de organización pertenecen: K-means (McQueen, 1967), K-medoid, Kerninghan-Lin, DK-Means, KNN, etc.

Desde el punto de vista de la posibilidad de la búsqueda metacontrolada los algoritmos se dividen en métodos descendentes y métodos competitivos, como ejemplo de ellos aparecen Simulated Annealing y los algoritmos genéticos.

Para dominios específicos aparecen los algoritmos de factorización de conceptos y los suffix tree clustering. Los algoritmos que aquí se exponen tienen diversas aplicaciones en los diversos métodos que se exponen. En el Capítulo I de esta investigación se aborda una gran cantidad de algoritmos que se usan en la agrupación de grupos en detalle.

#### 7.2.1.2.- Aplicaciones de los Algoritmos

La evaluación desarrollada por Lewis (Lewis, 1992) persigue estudiar las propiedades de partícula léxica y la indexación en clúster en tareas de categorización de texto, permitiendo evaluar sus posibilidades de agrupamiento de forma aislada para detectar problemas de interpretación. El autor demuestra la eficiencia de agrupamiento que se produce cuando se utiliza sólo una porción mímica de términos disponibles en la indexación. Lewis (Lewis, 1992) también demuestra que se consiguen menores niveles de efectividad sintáctica para la indexación de las palabras base. Aposta por una metodología que se compone de las siguientes fases: Colección de Términos mediante Categorización binaria, Representación textual, Clúster de Términos y Estrategia de Selección. Esta metodología facilita el análisis de los procesos de indexación de frases y palabras, dejando patente que los métodos de

agrupamiento tradicional no son los únicos para la representación textual. La aplicación de la metodología se basa en tres hipótesis de agrupamiento.

### 7.2.3.- Cálculo del VSM

(Vector Space Model) desarrollado en algunos trabajos de Salton. El Modelo Espacio Vectorial se fundamenta en la construcción de una matriz de términos contra términos que permite obtener una selección de los términos que subyacen en un corpus y construir con ellos un agrupamiento duro. El uso de esta métrica se ha visto utilizada en los trabajos de Cunha (Cunha et al., 2007) en el cual los autores utilizan VSM en un sistema que integra a CORTEX, un sistema de extracción de textos que utiliza un algoritmo de decisión con varias métricas de análisis basadas en (VSM), entre ellas la distancia Hamming, lo que le permite evaluar la calidad de los extractos a nivel oracional. CORTEX sigue la idea del bag of Word o sea la Bolsa de Palabras. Enertex es desarrollado para la Física Estadística. Utiliza métricas y algoritmos que se basan en el supuesto teórico que sustenta el tratamiento de un texto como una unidad interactiva de palabras, donde cada unidad de términos es afectada por el campo credo por otros. Estos sistemas se integran con Yate, sistema especializado en extracción de texto que posee una ontología que permite obtener todos los términos candidatos a través de un lexicón que combina varias técnicas léxico-semánticas de tratamiento de los términos. Al sistema se le añaden las reglas DISICO, un paquete de reglas de inferencia generados por D'Cunha (D'Cunha, 2006). Este sistema evalúa los corpus médicos, obteniendo resultados favorables, aunque es evidente que se necesitan mejoras experimentales para lograr resultados más específicos, que no sean exclusivamente precisión, recuperación y f-measure.

Otra variante del uso del VSM (Vector Space Model) y LRA (Análisis de Latencia Relacional) puede observarse en el trabajo de (Turney) el cual calcula la similaridad del corpus mediante el concurso de técnicas de latencia semántica. Este tipo de evaluación se centra en varios elementos para su evaluación entre los que se encuentran:

- Patrones derivados automáticamente del texto.
- Valores asociados a la descomposición del corpus, usada como clave de frecuencia.
- Uso de los sinónimos en la reformulación de los pares de trabajos.

Esta investigación permite utilizar sus resultados en actividades relativas a la desambiguación, traducción y extracción de textos.

En el terreno del VSM también aparece la aplicación metodológica de Rajan (Rajan et al., 2009) que sustenta la categorización de textos en tamil mediante redes neuronales artificiales. El método utiliza como centro base a VSM. Como primer paso la metodología selecciona el corpus de términos en tamil, usando stop word elimination y la asignación de pesos a las palabras indexadas. Luego se calcula el VSM utilizando dos distancias terminológicas: coseno y euclidiana. También aparece en una red neuronal que utiliza un algoritmo de decisión. El agrupamiento logrado con VSM es comparado con otro realizado por una Red Neuronal, siendo este superior al VSM por poseer una heurística que modela cognitivamente las formas posibles de clasificación.

En esta misma dirección también aparece la investigación doctoral desarrollada por Pinto Avedaño (Pinto Avedaño, 2008) en la cual se validan clústeres para evaluar la calidad de textos cortos. Este procedimiento matemático supone el uso de técnicas supervisadas y no supervisadas de evaluación de corpus. Pinto Avedaño (Pinto Avedaño, 2008) desarrolla una metodología coherente que parte de la posición que asumen hoy los textos libres en diversos repositorios. Para realizar este método se desarrollaron los siguientes procesos:

- **Fronteras del Dominio:** Factor necesario para delimitar un primer agrupamiento sobre las dimensiones del corpus.
- **Agrupamiento:** Considerando que la necesidad de las frecuencias de términos son imprescindibles para usar medidas de similaridad. Para resolver los problemas que ofrece las métricas de similaridad en textos cortos se apeló a las técnicas desarrolladas por Herdan (Herdan, 1960).

- **Balance de Clases:** Basada en la distribución homogénea de las clases y los clúster resultantes del agrupamiento.
- **Estilometría:** Este paso se detiene en el estilo de redacción del escritor del texto, ya que la existencia de textos con diversos estilos de producción (procedentes de investigadores o de documentalistas) pueden incidir en las métricas de evaluación y en la calidad de los clúster.
- **Estructura:** Se ha tenido en cuenta si el texto es científico, o sea estructurado o si es texto libre, pues cada uno presupone resultados diferentes para el agrupamiento y por tanto resultados diversos.

### 7.2.4.- Similaridad

La similaridad léxica es una de las medidas también utilizadas en Mihalcea (Mihalcea et al., 2006) en el proceso de evaluación de los textos. En el trabajo de Mihalcea (Mihalcea et al., 2006) se describe un método que toma como centro un corpus base y el conocimiento base. Para evaluar el corpus base se utilizó el coeficiente de Información Mutua, el cual permite evaluar casos de co-referencia y el coeficiente de Latencia Semántica. El conocimiento base es evaluado con el índice de Lesk. El resultado de la investigación patentiza que la técnica utilizada supera la tradicional evaluación realizada con vector base similarity.

Otras medidas que se aparecen en la literatura de las especialidad sobre este tema son Coeficiente de Jaccard (Frakes and Baeza, 1992), Tf-idf measure, Distancia Kullback-Leibler

En otra posición de la praxis aparece el trabajo de Roberts (Roberts et al., 2009) donde se describe un esquema para la construcción de un modelo de anotación para el dominio de la medicina, específicamente el de los pacientes con cáncer, del que se tomaron sus historias clínicas como corpus. El objeto pragmático de la técnica de evaluación es la construcción de un modelo de anotación semántico para ser utilizado en sistemas de minería de texto, para ello parten de una estrategia estructural el texto que define como claves de

anotación los siguientes puntos: investigación, condición, intervención y resultados.

En esta misma parcela Pineda (Pineda et al., 2002) desarrolló una metodología que permite transcribir y evaluar corpus acústicos, en la que se le aplican tres medidas de granularidad. A partir de la evaluación de esta metodología se han construido modelos acústicos y diccionarios fonéticos.

Una metodología que apareció hace algunos años, que demuestra la diversidad de estos estudios es Europar de Koehn (Koehn, n.d.), un proceder que evaluó la calidad de la traducción asistida por ordenador en textos generados por el Parlamento Europeo. Este corpus tenía como característica esencial, estar en 11 idiomas. Los métodos que se usaron en este examen fueron el análisis de las oraciones y el análisis de los documentos.

Otro algoritmo que se utiliza en estas técnicas es el de Waterman (Smith and Waterman, 1981) en el que se comparan los segmentos de todas las longitudes posibles, eligiendo el máxima similitud al obtener una secuencia óptima de alineaciones. Las aplicaciones más importantes utilizan como medida empírica 0.3.1 y dos para un espacio, copia y sustitución.

### 7.2.5.-Similitud Coseno

En este aspecto también aparece la similitud coseno. Es una medida basada en vectores donde las cadenas de caracteres de entradas se transforman en espacios vectoriales, siendo su cálculo el siguiente:

$$\cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

Otra distancia o similitud es la Distancia Jaro, es una medida donde el autor hace un análisis de las derivaciones posibles en la ortografía. Su cálculo se realiza de la siguiente forma:

$$d_j(s_1, s_2) = \frac{m}{3 \cdot |s_1|} + \frac{m}{3 \cdot |s_2|} + \frac{m - t}{3 \cdot m}$$

Siendo S1 y S2 las cadenas a comparar, $js_{1j}$ y $j_{s2j}$ y sus respectivas longitudes, m es el número de caracteres coincidentes, considerando sólo aquellos que no son mayores que el máx (js1j js2j, siendo 1 y t el número de trasposiciones realizadas por el ordenador, pero con diferentes caracteres.

### 7.2.6.- Índice de Kappa

En Raileanau (Raileanu et al., 1999) se utiliza el índice de Kappa para evaluar la calidad de la desambiguación de textos médicos. El objeto de esta investigación reside en la selección de términos ambiguos mediante anotación asistida. Los pasos que se realizaron para evaluar este método son los siguientes:

- **Selección de términos ambiguos:** de dos recursos terminológicos (Germanet, ULMS). En el caso de Germanet en primer momento se calcularon los valores de los términos que subyacen en el corpus., usando como métrica Tdf-ifs, en forma automática. Para ULMS no fue necesario usar ninguna técnica computacional ya que se parte del supuesto que este recurso tiene todos los términos relevantes y oficiales del lenguaje médico.
- **Anotación Guiada:** Para esto se calculó el índice de Kappa, que permitió evaluar la calidad de la desambiguación bajo sentido en los artículos médicos.

### 7.2.7.- Recobrado y Precisión

#### 7.2.7.1.- F- measure

En Arco (Arco, 2005, 2008) se describe F-score o F-measure. Es una medida de precisión de una prueba. Sus particularidades se basan en la precisión y la evidencia de la prueba, elementos que sirven para calcular la capacidad de

evaluación de un texto. Se valora el número de resultados correctos en la precisión y el recobrado, dividido el número de resultados devueltos entre r (es el número de resultados concretos) y dividido el número de resultados a obtener. Arco utiliza esta medida para evaluar el agrupamiento y la recuperación de su sistema SATEX.

$$F(C_i, C_j^*) = \frac{2 \cdot Precision(C_i, C_j^*) \cdot Recall(C_i, C_j^*)}{Precision(C_i, C_j^*) + Recall(C_i, C_j^*)},$$

### 7.2.7.2.- Rouge

En Fernández (Fernández et al., 2009) se utiliza Rouge para analizar textos en italiano. Es una medida muy utilizada en la evaluación de resúmenes y en los procesos de traducción automática. Es una técnica que puede combinarse metodológicamente para trabajar con n-grama. En Fernández (Fernández et al., 2009) se aplica una de las variantes de Rouge, denominada Rouge –N y Rouge –W. Estas medidas se explican detalladamente en el acápite 7.8.3.1 de la investigación.

### 7.2.8.- TDT (Detección de Tópicos y Localización)

La idea básica de la TDT se originó en 1996, cuando el Defense Advanced Research Projects Agency (DARPA) expresó la necesidad de utilizar tecnología para la detección de Tópicos de Noticias de forma automática. Los trabajos de Wayne (Wayne, n.d.) están formalizados desde las posibilidades de la IDT, auspiciada por el Proyecto DARPA. La misma centra sus basamentos en la detección de Tópicos como una postura para conformar y evaluar corpus de documentos. El sistema que utiliza el DARPA basa su funcionamiento en segmentación, detección, enlace y localización o Tracking.

Desde una decisión real, el software calcula el coste normalizado para cada tópico, demostrando la fortaleza de los algoritmos de tratamiento del texto en la construcción de umbrales. Debido a que los tópicos varían y a que la diversidad del tópico puede traer complicaciones el sistema pondera los resultados,

siendo una alternativa utilizada para mejorar la flexibilidad del sistema en que se emplee este método y su operatividad.

Una serie de experimentos con corpus para estudiar la generación de referencias para la expresión oral es descrita por Gupta y Stent (Gupta and Stent, n.d.), los que describen un experimento que evalúa automáticamente las expresiones a partir de reglas declaradas bajo algoritmos base, tratando que se adapten a otros o mejorando su flexibilidad. Los algoritmos evaluados con este método son: Baseline, Dale y Rater, Siddharthan y Copestake, Partner-Specific Adaptation y Partner-Specific Adaptation Variant. La evaluación consiste en observar las modificaciones que ocurren cuando estos algoritmos son implementados en el Corpus Coconut, determinando su capacidad de generar non-frases. Los indicadores que mide esta metodología son bigramas, uningramas y randon.

### 7.2.9.- Evaluación de Herramientas y Software

En Carroll (Carroll et al., 1999) se define una metodología que permite evaluar un parser que ha de etiquetar léxico-semánticamente un texto corto. El esquema de relaciones a través de las que se establecen las marcas es establecido a través de cabezales y dependencias. Las medidas utilizadas en esta investigación son precisión, recobrado y f score. Con esta investigación se muestra una nueva forma en el desarrollo de las investigaciones sobre el etiquetamiento automático de los textos.

### 7.2.10.- Reglas de Aprendizaje

En este apartado aparecen como muy relevantes los trabajos de Bouillon (Bouillon et al., n.d.) en los que se evalúa la calidad del corpus mediante reglas de aprendizaje. A base de este sistema es la existencia de pares léxicos nombre-verbo. Para construir las heurísticas este sistema se basa en normas de expresión semántica, recogidas de textos etiquetados semánticamente. Esta variante utilizada en este sistema supera los resultados alcanzados con otros

sistemas sustentados en posturas estadísticas y sintácticas, tradicionalmente utilizadas en estas investigaciones.

### 7.2.11.- Métodos Estadísticos

#### 7.2.11.1.- Covarianza

En Kaalep y Veskis (Jaan Kaalep and Veskis, 2007)se ha desarrollado una evaluación de corpus, en inglés y en estonio utilizando métodos automáticos, a partir de colecciones creadas independientemente. Para valorar la similaridad de los textos se utilizó la covarianza de los resultados obtenidos cuando el proceso fue hecho bajo supervisión y automáticamente.

### 7.2.12.- Cartografía Documental

Esta clasificación pone de manifiesto aquellas evaluaciones que analizan los cartogramas generados con mapas de conocimiento o estructuras conceptuales de dominio.

Los trabajos de Shams y Elsayed (Shams and Elsayed, 2008) reflejan como se construye un Corpus Base para la evaluación TKM (Text Knowledge Mapping Prototype). A partir de MORPHIX, una herramienta de análisis léxico se evalúan los componentes léxicos del corpus en el tema de la ingeniería eléctrica. Esta metodología facilita la construcción de mapas documentales. Las fases del modelo de evaluación son dos: Evaluación de los componentes léxicos del Vocabulario, Evaluación del modelo de conocimiento. El primero se centra en el análisis de los coordinativos, los pronombres, los adjetivos, las preposiciones, los verbos y los adverbios. Con estos elementos se evalúa las palabras en el corpus y la cobertura del vocabulario. El segundo se basa exclusivamente en (análisis de las relaciones semánticas de la ontología y en el corpus, análisis retórico, Análisis de Discurso y Análisis de Tópicos en el Discurso.

### 7.2.13.- Evaluación mediante Métodos Empíricos

En este apartado aparecen los trabajos de Onciniz-Martínez (Onciniz-Martínez, 2009) los cuales están centrados en el estudio de la presencia de anglicismos en el español contemporáneo. Desde el punto de vista metodológico esta investigación aporta una base empírica para el reconocimiento de la fraseología inglesa y su influencia en el español. Para el estudio se usó el corpus CORDE y el CREA (corpus de la Real Academia Española). También se utilizan dos corpus de inglés, el BNC y el COCA con el objetivo de hacer la contrastación léxica.

En Parsons (Parsons et al., 2009) se presenta un estudio que parte de las posibilidades del estudio empírico de las caligrafía en un texto generado por humanos. El estudio mide la similaridad a partir de los valores que facilita la herramienta de análisis desarrollada por Kaligarrifff 2001(Kilgariff, 2001) con ($X^2$). Los resultados del cálculo de ($X^2$) son corroborados con medidas humanas. Los corpus que se evalúan son académicos, mensajes y corpus libres. El experimento cuenta de tres partes: la valoración humana, la valoración de ($X^2$) y la comparación de la similaridad de los obtenidos en ($X^2$) y las valoraciones humanas.

Todas estas metodologías son hijas de sus aplicaciones, no cabe duda, sin embargo hay una visión muy clara los resultados metodológicos en este campo adquieren más confiabilidad si en el experimento y en la metodología se miden indicadores que parten del conocimiento que aportan los humanos y la capacidad de análisis de las máquinas.

### *7.2.14.- Metodología de Evaluación del Corpus de Puertoterm, propuesta de desarrollo*

Se ha desarrollado una metodología que aúna los criterios netamente matemáticos y aglomerativos con criterios netamente cognitivos y lingüísticos. Es por ello que esta metodología se parte de técnicas de selección de términos y se culmina con Técnicas Cognitivas de Evaluación de textos. La metodología

debido a estas necesidades se divide en dos partes: Técnicas de Selección de términos y Técnicas de Evaluación de Corpus. Las primeras son eminentemente matemáticas y se implementan por medio de FOXCORP, herramienta diseñada adhoc por de Domínguez (DOMÍGUEZ, 2011) y las segundas se aplican a observaciones de los expertos en el corpus del proyecto Puertoterm, dichas observaciones son estudiadas mediante el uso de la técnica Chi cuadrada.

Los corpus que han servido sustento para esta investigación poseen características muy disímiles en lo referente a nivel discursivo, prestigio de sus fuentes y la tipología documental. Esta situación hace que las estrategias de trabajo y análisis sean más complejas, debido a la disparidad en la concepción del corpus.

En el proyecto Puertoterm está desarrollado sobre dos corpus: Uno en inglés y otro en español. El corpus inglés es eminentemente especializado (más de 4000) registros están en este orden, (818 son semi-especializados) y solo 43 son artículos divulgativos. El corpus español es, en esencia, divulgativo y en menor grado especializado.

Para el desarrollo de este artículo se han utilizado varias técnicas de investigación. En el terreno de la lexicometría se utilizó el software FOXCORP desarrollado por (DOMÍGUEZ, 2011) con el fin de etiquetar automáticamente los corpus, debido a que la mayoría de las herramientas utilizadas a tales efectos solo estaban desarrolladas para textos en inglés. Con FOXCORP se etiquetó un 25 % de cada uno de los corpus determinado la muestra por métodos aleatorios y con TEXMIX (Domínguez, 2011) se construyeron las correspondencias terminológicas que facilitaron reducir la dimensionalidad. Para buscar representatividad en los corpus se aplicaron técnicas aleatorias de selección, lo que arrojó que debían analizarse 1115 textos en español y 1678 en inglés.

### 7.2.14.1.- Técnicas de Selección de Términos

El algoritmo de trabajo desarrollado para el análisis de los variables de esta investigación ha sido descrito y desarrollado por Arco (Arco, 2005) e implementado en FOXCORP (Domínguez, 2011). Esta herramienta facilita el uso de medidas de estudio del léxico y medidas de validación de la calidad del agrupamiento, las que al ser combinadas dan un alto grado de fidelidad a la valoración de cada corpus. El procedimiento consiste en lo siguiente:

- **Transformación de los corpus**: hasta obtener cada fichero en forma de token de palabras.
- **Obtener una Representación VSM**: Es conocido que VSM parte de la construcción de vectores generados a partir de documentos en los que existe determinada ocurrencia de términos. En Arco (Arco, 2007) se construye una matriz donde los términos indexados se encuentran en sus respectivas filas y los documentos de ambos corpus en columnas, en las que cada celda denota la frecuencia en que aparece cada vocablo en un documento. Arco (Arco, 2007) también construye la matriz considerando las columnas cada una de las oraciones o párrafos que conforman los documentos del corpus, por tanto, cada celda corresponde a la frecuencia de aparición de los términos en la oración o párrafo, respectivamente. Sea $d \in D$ un documento textual. La representación de $d$ es el vector documento $d=\rho(d)=(w_1,\ldots,w_m)^T \in R=\mathbf{R}_{m+}$, donde cada dimensión corresponde a un término en la colección de documentos y $w_i$ denota el peso del $i$-ésimo término. El conjunto de esos $m$ términos indexados, $V=\{t_1,\ldots,t_m\}$, es referido como el vocabulario.
- **Extracción de términos:** Tomando como base los vectores de términos obtenidos en el paso anterior, a partir de la Representación VSM se produce una secuencia de términos indexados contra documentos que generan como productos un token.
- **Normalización de la matriz:** Según Arco (Arco, 2005) Es más eficiente normalizar los vectores pesados antes de involucrar las operaciones en

el cálculo de la similitud (e.g. al agrupar documentos). Arco (Arco, 2008) pasa por alto la variedad de longitudes de documentos normalizando la matriz a partir de la división de cada frecuencia absoluta de aparición de un término t en un documento d, por la suma de los componentes del vector pesado del documento.

- **Calcular las medidas de Calidad de los términos:** Para ello se realizan los siguientes cálculos:

    - Frecuencia en el corpus
    - Rango en el Corpus
    - Frecuencia como palabras Clave
    - Rango como palabras
    - **Reducción de la dimensionalidad**: Se apela al procedimiento eliminación de palabras vacías mediante un diccionario.
    - **Calcular las operaciones de Similitud:** Las operaciones de similitud se hacen a partir de operaciones estadísticas simples hasta lograr otras más complejas. Las variables que se usaron en esta evaluación fueron las siguientes:

1. **Palabras temáticas importantes:** esta técnica de análisis estadístico de la estructura del texto opera mediante la identificación de términos clave dentro de los documentos.
2. **Umbral de frecuencia de términos y Ley de Zip.** Una heurística de selección muy simple que elimina todos los términos cuyas frecuencias son superiores a un umbral predefinido o inferiores a un umbral predefinido. A partir de observaciones hechas por Lunh (1958), esta medida es tomado como un indicador de significación. Por tanto, la frecuencia de ocurrencias de términos es una medida apropiada de la significación de los términos (Lanquillon, 2002). Así, términos que raramente aparecen en una colección de documentos tendrán poco poder discriminante y pueden ser eliminados (Risberjen, 1979). En contraste, términos con frecuencia de aparición alta se asumen que son comunes y que tampoco tienen poder discriminante (Anexo del 23-25).

3. **Umbral de frecuencia de documentos.** Utilizado para obtener la frecuencia en de aquellos términos que aparecen en grupos textuales para determinar si el corpus tiene validez léxica o no. Además, usar la ocurrencia de términos infrecuentes no es confiable estadísticamente. Al eliminar estos términos se mantiene el poder discriminante y se mejora la efectividad del agrupamiento y clasificación textual.
4. **Razón de señal a ruido (Signal-to-Noise Ratio).** Determina el ruido que puede crear un término en un corpus dado. Tomando en consideración la teoría de la información, la razón de señal a ruido de un término particular mide el poder discriminante que transmite ese término, por tanto los términos con grandes valores son preferidos. La evaluación del ruido del entorno se basa en la entropía (Salton and Macgillm, 1983). Ésta es una variable aleatoria discreta *X* sobre *c* valores diferentes con probabilidades *pi*, *i*=1,…,*c*, definida por:

$$\text{Entropía}(X) = -\sum_{i=1}^{c} p_i \log p_i$$

Así, la entropía puede ser evaluada como la cantidad de información que esperamos recibir sobre el promedio cuando observamos una variable aleatoria particular. La entropía es cero cuando una salida de la variable aleatoria ocurre con certeza, i.e. $p_i$=1 para un *i* arbitrario y $p_j$=0 para $j \neq i$. Mientras más uniforme sea una distribución, mayor es su entropía. Así, la entropía alcanza su máximo valor log *c* si todas las salidas de *X* son igualmente probables. Cuando un término *t* está concentrado en sólo pocos documentos, se puede calcular el Ruido (*t*) (Salton and Macgillm, 1983) como la entropía de la distribución de probabilidad del término *t* entre los documentos:

$$\text{Ruido}(t) = -\sum_{j=1}^{n} \text{P}(d_j,t) \log \text{P}(d_j,t), \quad \text{donde} \quad \text{P}(d_j,t) = \frac{tf_{d_j}(t)}{tf(t)}$$

es decir, la probabilidad que un documento $d_j$ y un término $t$ co-ocurran. El rango de la función ruido es [0,log $n$]. El ruido es cero cuando el término $t$ aparece sólo en un documento y toma su valor máximo log $n$ si el término $t$ ocurre con la misma frecuencia en todos los documentos. La razón señal a ruido se expresa como la diferencia de esos logaritmos SNR($t$)=log $tf(t)$ – Ruido($t$). Se utiliza la entropía, según Lochbaum y Streeter en 1989 (citados por Leiva, 2009, et.al.) como una medida para el cálculo de la importancia de las palabras:

$$\text{Entropía}(t) = 1 + \frac{1}{\ln(n)} \sum_{i=1}^{n} p_i(t) \cdot \ln(p_i(t)) \quad \text{donde } p_i(t) = \frac{tf_{d_i}(t)}{\sum_{j=1}^{n} tf_{d_j}(t)}$$

Empíricamente se consideran relevantes las palabras que tienen una alta entropía, dentro de aquellas que tienen una alta frecuencia de aparición (i.e. preferimos seleccionar aquellas palabras que tienen una entropía alta desde un conjunto de palabras que son igualmente frecuentes).

5. **Calidad de términos.** Consiste en determinar la cantidad de términos relevantes que posee cada corpus (Anex0 23-25).
6. **Skewness y Kurtosis:** Mide la capacidad de un término en este caso la de parcialidad o sea su capacidad para aparecer en determinado corpus (Anexo 23-25).
7. **Overall Similarity.** Utilizada para calcular la similitud del agrupamiento y observar cual es la similitud de los grupos de palabras y su valor homogéneo (Anexo 23-25).
8. **F-Measure:** Mide la capacidad del los grupos de palabras generados a partir del texto para ser recuperados (Anexo 23-25.

### 7.2.15.- Técnicas Cognitivas de Evaluación de Corpus

Entre las técnicas cognitivas para el análisis de estos textos está el análisis por los humanos, sustentados la teoría de Hernández (Hernández, 2007) en la que se expone que los valores del texto fuente deben ser evaluados por los humanos mediante observación. Para esta evaluación se tiene en cuenta la desviación estándar a partir de las observaciones de 2 expertos en el tema, se ha utilizando la fórmula de la similaridad para calcular en cada caso los valores que se necesitan. Las medidas que se utilizarán para este paso del artículo serán las siguientes:

**Similaridad Estructural:** Mide la presencia en diversos pare de documentos del corpus de Puertorem, teniendo en cuenta la estructura retórica, entidades, formas lingüísticas, párrafos.

**Similaridad en el Uso del Lenguaje:** Evalúa el léxico, su estructura sintáctica, gramatical, tipos de oraciones, uso de signos de puntuación, posición de los verbos.

**Similaridad Estructural en el Uso de las Imágenes:** Evalúa la presencia en los textos de Puertoterm de imágenes, formato de imagen, claridad, trasparencia y correspondencia de la imagen con el texto.

**Similiaridad de Contenido:** Se basa en la postura del contenido y su tratamiento temático.

**Frecuencia de las Categorías Gramaticales:** Existencia de verbos, adverbios, homónimos, merónimos, hipónimos, homónimos.

## *7.3.- Resultados de la Evaluación del Corpus*

### 7.3.1.-Umbral de Frecuencia de Términos

Teniendo en cuenta que n(t) es el número de documento en los cuales el término X aparece al menos una vez, es posible utilizar como medida de

selección la exclusión de aquellos vocablos sea menor que el umbral de 5, ya que términos que ocurren en sólo muy pocos documentos son vocablos cuya información no es válida para denotar grupos textuales. Esto crea ruidos y falsas estructuras de comunicación. La supresión de estos términos posibilita la capacidad de discriminar y mejora la calidad de los resultados del agrupamiento y el etiquetaje de los grupos (Ver tabla 59).

| Umbral de Frecuencia de Términos | Corp. Español | Corp. Inglés |
|---|---|---|
| | 45 | 27 |

Tabla 59. Frecuencia de Términos

Con esta medida se determina la probabilidad de encontrar documentos en el corpus que no refleje al menos 5 documentos, lo que permite eliminar aquellos documentos que no tienen valor para el desarrollo del corpus terminológico.

El resultado de esta medida indica que hay que eliminar muy pocos documentos en ambos corp., sin embargo en el texto inglés la frecuencia de términos que aparecen en el corpus, es menor, dada la existencia de diversas estructuras lingüísticas que hacen que el español sea más rico terminológicamente.

### 7.3.1.2.- Calidad de Términos:

Para realizar este cálculo se muestran dos medidas que son utilizadas para medir la calidad de los términos y por tanto permiten la reducción de la dimensionalidad a partir de la selección de aquellos términos relevantes (Tabla 60).

$$q_0(t) = \sum_{j=1}^{n} \left(tf_{d_j}(t)\right)^2 - \frac{1}{n}\left[\sum_{j=1}^{n} tf_{d_j}(t)\right]^2 \quad (1.9) \qquad q_1(t) = \sum_{j=1}^{n_1} \left(tf_{d_j}(t)\right)^2 - \frac{1}{n_1}\left[\sum_{j=1}^{n_1} tf_{d_j}(t)\right]^2$$

| Calidad de Términos | Corp. Español | Corp. Inglés |
|---|---|---|
| | 32 | 48 |

Tabla 60. Calidad de los Términos

En cuanto a calidad de los términos, es el corpus español superior al corpus inglés. Las medidas de calidad del corpus inglés indican que es necesario eliminar el 0.32 % de los términos del corpus español y el 0.48 % del corpus inglés cuando se reduzca la dimensionalidad, es decir cuando se eliminen las palabras que posean poca relevancia semántica, connotacional y pragmática.

### 7.3.1.3.- Umbral de Frecuencia:

El método de Lunh (1958) sobre la aparición de términos en un texto es tomado como un índice de validez y un indicador de significación. La ocurrencia de términos constituye un estándar apropiado para valorar la significación de los términos. Debido a esto términos que posen baja frecuencia de aparición en la colección son tomados como índices de baja capacidad discriminante por tanto pueden eliminarse, evitando así ruidos en las colecciones. Sin embargo, aquellos vocablos cuya frecuencia de aparición alta son vistos como comunes o generales, por tanto su poder discriminatorio es bajo. En esta investigación el umbral de frecuencia de los términos se comporta de la siguiente forma (ver tabla 61).

| Frecuencia de Términos | Corp. Español | Corp. Inglés |
|---|---|---|
| | 56 | 42 |

Tabla 61. Frecuencia de Términos

Es evidente que los textos españoles son menos ricos en existencia de términos relevantes, debido a la existencia de muchos sinónimos, parónimos, merónimos y homónimos. El corpus Inglés al ser más especializado, presenta

mayor frecuencia de aparición de términos. La especialización que presenta el texto inglés hace que términos más ricos y cercanos al terreno de la Ingeniería de Puertos y Costas aparezcan en él, pues este dominio lingüístico tiene sus bases especializadas en esta lengua. A continuación se exponen las medidas de validación de los agrupamientos

#### 7.3.1.4.- F-measure:

F-Measure, es una métrica que combina las ideas de precisión y recuperación de información, se trata cada clúster como si este fuera el resultado de una consulta y cada clase como si esta fuera el conjunto de documentos deseados para una consulta. Así, se calcula precisión y recobrado de los clústers para cada clase dada. Más específicamente, para el clúster j y la clase i

$$F(i,j) = \frac{2 \cdot recall(i,j) \cdot precision(i,j)}{recall(i,j) + precision(i,j)}$$

donde nij es el número de miembros de la clase i en el clúster j, nj es el número de miembros del clúster j y nj es el número de miembros de la clase i (Ver tabla 62).

| Frecuencia de Términos | Recall | Precisión |
|---|---|---|
| C.Inglés | 0.80 | 0.70 |
| C.Español | 0.90 | 0.50 |

Tabla 62. Recobrado y Precisión

#### 7.3.1.5.- Skewness y kurtosis

Son medidas estadísticas que refieren una deformación en las distribuciones. Su validez instrumental facilita conocer los términos que están solapados o parcializados. Parcialidad de un término t se define como:

$$P(t)=w_1 \cdot \text{Skewness}(t)+w_2 \cdot \text{Kurtosis}(t)$$

donde w1 y w2 son pesos positivos para Skewness y Kurtosis, definidas a continuación.

$$\text{Skewness}(t)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(tf_{d_i}(t)-\frac{tf(t)}{n}\right)^3}{s^3} \quad (1.11) \qquad \text{Kurtosis}(t)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(tf_{d_i}(t)-\frac{tf(t)}{n}\right)^4}{s^4}-3$$

donde s es la desviación estándar de la ocurrencia del término t en el Corpus de documentos. Valores altos de Skewness (t) y Kurtosis (t) indican que el término t es más general en el corpus de textos y viceversa (Tabla 63).

| Frecuencia de Términos | Skewness | Kurtosis |
|---|---|---|
| C.Inglés | 0.56 | 0.50 |
| C.Español | 0.96 | 0.78 |

Tabla. 63. Skewness y Kurtosis

### 7.3.1.6. - Overall Similarity

Como nos encontramos con un corpus etiquetado a nivel de palabras, sin la presencia de clases ni etiquetas de clases se ha decidido valorar la unidad interna de los clúster que forman ambos corpus, usando para ello la similitud pesada interna de los dos clúster de los corpus por separado (Ver Anexo 23-25).

$$OverallSimilarity=\frac{1}{|S|^2}\sum_{\substack{d\in S\\ d'\in S}}\text{distance}(d',d)$$

donde ISI es el número de documentos que pertenecen al clúster a evaluar., algunos autores utilizan el cociente Coseno para calcular la distancia entre los vectores, en esta investigación se decidió desarrollar la similaridad para

reconocer la calidad de los grupos de palabras formadas con el corpus (Ver tabla 64).

| Indicador | Overall Similarity |
|---|---|
| C.Inglés | 0.56 |
| C.Español | 0. 96 |

Tabla. 64. Overall Similarity

### 7.3.1.7.- Razón de señal a ruido (Signal-to-Noise Ratio)

Es la postura que posee un término particular al producir ruido en la comunicación es por ello que la base de su cálculo se sustenta en la entropía. Es la entropía la distribución de la probabilidad del término t entre en determinado documento (Ver tabla 65 y Anexos del 23 al 25).

$$\text{Entropía}(t) = 1 + \frac{1}{\ln(n)} \sum_{i=1}^{n} p_i(t) \cdot \ln(p_i(t)) \quad \text{donde } p_i(t) = \frac{tf_{d_i}(t)}{\sum_{j=1}^{n} tf_{d_j}(t)}$$

| Entropía | C.Español | C.Inglés |
|---|---|---|
| | 0.47 | 0.13 |

Tabla 65. Entropía

El ruido es más probable en los términos del corpus español que en el de inglés, debido a su gran cantidad de términos con estructura discursiva con carácter de sinonimia.

#### 7.3.1.7.1.- Ténicas de Evaluación de Corpus

El resultado de la evaluación de la similaridad basado en el en estudio de dos expertos valorado a través de Chi cuadrado para el estudio de 1115 textos en español y 1678 en Inglés con 789 y 894 palabras obtenidos a partir del los clúster. Los participantes evaluaron la similaridad estructural, la similaridad léxica y la similaridad de contenido comparándolos con $X^{2,}$ una medida

estadística que facilitó determinar con claridad los resultados de la valoración de los jueces y la aplicación de otras técnicas estadísticas. Los resultados de la evaluación del corpus español y el inglés evidencian las distorsiones en el proceso de evaluación. Se aprecia que el corpus inglés supera en estructura, léxico y contenido al corpus español. Los valores que le otorgan los jueces al corpus inglés difieren en ambos jueces, sin embargo el juez A le da al corpus inglés el mismo valor en el resultado final. Los resultados que aparecen en el estudio del $X^2$ pueden corroborarse si se observan las frecuencias de las palabras en ambos corpus. Los participantes en el experimento alcanzan su mayor nivel de coincidencia cuando valoran el la similaridad léxica, dándole al corpus inglés los valores más próximos en este parámetro A (68 %) y B (67 %). La riqueza terminológica del corpus español hace que las valoraciones sobre él sean más bajas cuando es valorado por un experto. Por su parte la estructura discursiva del texto inglés le otorga mayor valor estructural. En el aspecto de la similaridad estructural, el corpus inglés es más consistente en estructura, esto debe permitir la construcción de sistemas basados en la retórica del texto, pues los valores declarados por los expertos A y B son de 65 y 64 %, sin embargo el corpus español al ser más divulgativo es menos estructurado, y la variación de criterios en este indicador por los expertos en muy grande (0.40 %) y (0.46). Las imágenes son los elementos que más valoración reciben en el corpus español, siendo este el aspecto en que el corpus español supera al inglés. En la similitud del lenguaje se aprecian mejores coincidencias 0.68 y 0.67 para el corpus, debido a que el léxico inglés posee menos diversidad de palabras. El corpus inglés posee mayor nivel de similaridad con valores para A y B de 0.65 y 0.64 % (Ver Tablas 66-69).

## Coeficiente $X^2$ para Corpus Inglés

| Métricas | Cat.Iguales | Cate.Diferentes | Resultados |
|---|---|---|---|
| Similaridad de Contenido | A B | A B | A B |
| | 0.48 0.40 | 0.59 0.60 | 0.64 0.59 |
| Similaridad Estructural | 0.46 0.44 | 0.37 0.49 | 0.64 0.65 |
| Similaridad Lenguaje | 0.53 0.52 | 0.28 0.47 | 0.68 0.67 |
| Similaridad/imag. | 0.53 0.52 | 0.28 0.37 | 0.67 0.68 |
| Overall Similarity | 0.49 0.45 | 0.41 0.48 | 0.65 0.64 |

Tabla 66. Métricas de Evaluación Corpus Inglés

## Coeficiente $X^2$ para Corpus Español

| Métricas | Cat.Iguales | Cate.Diferentes | Resultados |
|---|---|---|---|
| Similiradidad de Contenido | A B | A B | A B |
| | 0.45 0.47 | 0.56 0.57 | 0.61 0.56 |
| Similaridad Estructural | 0.43 0.41 | 0.31 0.46 | 0.42 0.62 |
| Similaridad Lenguaje | 0.50 0.49 | 0.25 0.44 | 0.50 0.46 |
| Similaridad/imag. | 0.67 0.66 | 0.61 0.63 | 0.56 0.58 |
| Overall Similarity | 0.51 0.50 | 0.43 0.52 | 0.56 0.66 |

Tabla 67. Métrica de Evaluación Corpus Español

## Frecuencia de Términos

| Corpus | verbos | adverbios | homónimos | merónimos | hipónimos | Total |
|---|---|---|---|---|---|---|
| C.Español | 234 | 56 | 267 | 98 | 239 | 894 |
| C.Inglés | 145 | 29 | 178 | 270 | 167 | 789 |

Tabla 68. Frecuencia de Términos

**Tipología Discursiva**

| Corpus | Artículos Científicos | Artículos Léxicos | Materiales Divulgativos | Total |
|---|---|---|---|---|
| C.Español | 272 | 78 | 544 | 894 |
| C.Inglés | 561 | 200 | 28 | 789 |

Tabla 69. Tipología Discursiva

### *7.3.1.8.- Resultados de la Evaluación de Corpus*

La metodología propuesta muestra una nueva vía donde se conjugan las técnicas de clúster con modelos cognitivos de análisis. Con esta metodología se reafirman las posibilidades de VSM para aplicar técnicas de calidad de términos, lo que le confiere robustez a esta técnica. Los resultados que se obtienen mediante la aplicación de las Técnicas de Evaluación de Corpus y las Técnicas de Selección de Términos permiten obtener una visión global de la calidad de un corpus con vistas a su selección para la confección de recursos terminológicos como ontologías, taxonomías y diccionarios temáticos. El corpus inglés es superior al corpus español, pues su nivel léxico, la calidad de sus términos, su estructura, y su contenido arrojan valores superiores al del español.

## *7.4. - Evaluación de Ontosatcol*

En la representación del conocimiento es cada vez más común el uso de las ontologías como herramientas que permitan relacionar los conceptos que aparecen dentro de una materia. Estas especificaciones formales cuentan con varios mecanismos y fases para ser construidas, así como múltiples proyectos y publicaciones que avalan la validez de todas las metodologías propuestas. En general, de todas ellas se puede concluir lo mismo: no hay un único camino para construir una ontología, pero el producto resultante debe ser consecuente con los orígenes que lo propiciaron.

Al mismo tiempo, y de casi forma natural, surge la necesidad de evaluar lo realizado. La mayoría de metodologías empleadas para la creación de las ontologías también incluyen un apartado dedicado a la evaluación del resultado. Pero, en casi todos los casos, esta evaluación está excesivamente mediatizada por el sistema empleado en la formalización (por ejemplo, methontology de Fernández López (Fernández-López et al., 1997) o, directamente, no informa sobre cómo debe llevarse a cabo, pero sí indica la necesidad de ello como sucede con Uschold (Uschold and King, 1995), o es demasiado subjetiva (caso, por ejemplo, de Grunninger y Fox (Grunninger and Fox, 1995).

Por otra parte, hay que recordar que una de las claves del éxito de las ontologías se encuentra en la posibilidad que ofrecen para su reutilización. Es decir, emplear total o parcialmente otra/s ontología/s para adaptarla/s a la nueva que se está creando. Si bien es cierto que cada ontología responde a una manera diferente de contemplar o comprender el estado de un conocimiento concreto, no es menos evidente que los conceptos, conceptos son, y siempre es posible emplear sólo elementos determinados en función de las necesidades (capturar la taxonomía principal o tratar la ontología origen únicamente como base terminológica, obviando otras propiedades, axiomas, instancias, etc., para la ontología destino son las dos soluciones más comunes). Sea cual sea la opción, siempre es necesario realizar una evaluación de las ontologías candidatas con el fin de averiguar cuál se adapta mejor a los requerimientos.

En la actualidad existen diversos mecanismos o métodos que permiten evaluar ontologías entre ellos se encuentran los de Evermam, Fang y Staab (Evermann and Fang, 2010, Staab, 2004). Los trabajos de Gómez-Pérez (Gómez Pérez, 1994) apuntan hacia la existencia de varias formas de evaluación de recursos ontológicos desde el punto de vista industrial, cognitivo y con el empleo de técnicas matemáticas generalmente conocidas como métricas. Varios proyectos de nivel ontológico (Yu et al., 2009) han sido objeto de evaluación mediante el empleo de diversas técnicas y presupuestos teóricos (ANSI, 1978),

encaminados a evaluar esencialmente la expresión, la precisión, los problemas de diseño y las inconsistencias semánticas.

Una de las aplicaciones que se han desarrollado en el campo de la organización del conocimiento por medio de ontologías es Satcol de Leiva y otros (Leiva et al., 2009). Se trata de una aplicación que permite realizar procesos de minería de texto con los documentos de un corpus concreto, facilitando la creación de relaciones de acuerdo a la estructura gramatical y conceptual que presenten. Este programa emplea a un conjunto de agentes que se encargan de la modelación, el análisis documental y la simulación del comportamiento de resumidores. Para llevar a cabo ese trabajo se emplea Onto-Satcol, ontología formada por una estrategia cognitiva y otra semántica que sirve para expresar las relaciones semánticas, interpersonales y de guía en el trabajo de los agentes.

Al ser Satcol una herramienta para representar información en el terreno de la Ingeniería de Puertos y Costas, provoca que la ontología resultante sea "diferente" al resto en cuanto a diseño y desarrollo. El hecho de no contar esta disciplina con diccionarios especializados y la ausencia de una teoría específica debido a la juventud de esta parcela del conocimiento, obligan a buscar métodos diferentes para la evaluación de la ontología. Onto-Satcol no posee los cánones clásicos (a nivel conceptual) de diseño ontológico como otros modelos de ontologías. Su estructura responde a la desambiguación del conocimiento generado en disímiles terrenos, entre los que se encuentran: medio ambiente, costas, puertos, ingeniería de caminos y canales.

Con Onto-Satcol se pretendió solucionar el problema de integrar los conocimientos que eran necesarios para la ingeniería de Puertos y Costas y contribuir así a la construcción de un sistema sumista que se apoyara en la semántica de la ontología para extraer, desambiguar y recuperar información.

Este segmento de la investigación tiene como objetivo mostrar los pasos seguidos en la evaluación de la ontología Onto-Satcol, empleando para ello algunas métricas y principios específicos desarrollados pensando en las

características concretas de este sistema. De esa forma, además de medir su eficacia, estaremos mostrando una metodología diferente para evaluar ontologías que se puede unir a las comentadas anteriormente.

### 7.4.1. - Los métodos de evaluación de ontologías

Los modelos de evaluación de ontologías poseen diversas aplicaciones, de acuerdo a la ontología en cuestión que se esté analizando. Las formas de evaluar los sistemas ontológicos se han basado esencialmente en la semántica externa de las ontologías, la precisión de los modelos cognitivos, la estructura del conocimiento representado así como en medidas basadas en los criterios de exhaustividad y precisión Kent (Kent, 1955)clásicos de en la recuperación de información.

El autor ha decidido agrupar los métodos esenciales de evaluación para, después, pasar a describir el modelo que servirá de base a la ontología Onto-Satcol. Teniendo en cuenta que el objetivo de este texto no es el de establecer una clasificación de sistemas y métodos, algo que por otra parte ya ha sido propuesto en otros trabajos que tratan ese tema de manera monográfica, como los de Brank (Brank et al., 2005), Obrst (Obrst et al., 2007) o Hartmann (Hartmann et al., 2005), no se seguirá un esquema lingüístico o semántico en su listado, ni por supuesto recorrido exhaustivo, sino que se desarrollará sobre parámetros de practicidad.

#### 7.4.1.1.- Empleo de criterios clásicos en la recuperación de información

La primera variante utiliza dos métricas estándar en la recuperación de información para valorar la ontología: precisión (proporción de material relevante recuperado –bien por medio de clases o subclases, bien por medio de objetos asociados a éstas, etc.- del total de objetos recuperados) y exhaustividad (capacidad de la ontología para recuperar objetos –clases, subclases, documentos, etc.- relevantes del total de objetos relevantes almacenados en la ontología). En la mayoría de trabajos que han aplicado estos criterios se ha contemplado tanto la dimensión semántica, inherente a

cualquier ontología, como la más pura vertiente de recuperación de información.

La aportación más interesante en este tipo de trabajos la presenta Zhang (Zhang et al., 2008.), quien defiende una relación puramente semántica del concepto de precisión, dentro de la posición y evaluación de mapas ontológicos. Para evaluar dichos mapas, él y sus colaboradores comparan los criterios puramente semánticos con otros valores normalizados y simétricos.

Otro aspecto que caracteriza las evaluaciones realizadas a partir de estos criterios es la multiplicidad de variantes que, tanto precisión como exhaustividad, pueden aportar, y que son comentados en diversas contribuciones, en la línea defendida por Ramos (Ramos et al., 2009) que presenta una variante para calcular la exhaustividad y la precisión basada específicamente en las capacidad semántica de la ontología y su operatividad en determinado corpus. Sin embargo, las fórmulas con las que trabaja son las típicas de estos dos criterios, y los valores resultantes deberían ser completados con la medida de F, que sirve para corregir el error de la distancia que se produce en aquellos casos en los que la exhaustividad y la precisión se compensan.

Un sistema que emplea un mecanismo parecido, y que presenta ese mismo inconveniente, es Evalexon (Spyns, 2005), que trabaja de manera conjunta con la ontología y el corpus sobre el que ésta se ha generado. Emplea diversas técnicas de minería de texto para realizar una evaluación más conceptual que lingüística, ya que se preocupa más de cómo de apropiados son los términos escogidos para representar conceptos, en función de la frecuencia con la que esos términos aparecen tanto en la ontología como en el corpus.

El principal problema que presentan estos métodos de evaluación es la necesidad de un entrenamiento previo de los observadores, para que puedan valorar las conclusiones de manera objetiva. En muchas ocasiones, los resultados no son los más factibles, debido a que tanto precisión como exhaustividad son dos cuestiones que se deben interpretar desde el punto de

vista de la persona que interactúa con el sistema. El no contar con un criterio formal (expresado en cantidades) o una visión semántica clara (expresada por medio de la posición que existe entre el objeto que se obtiene y lo que el usuario necesita) son los principales hándicaps del empleo exclusivo de este método.

La segunda variante emplea las posibilidades que ofrece el modelo de espacio vectorial Salton (Salton et al., 1975) dentro de una ontología y un corpus de texto. Este modelo de evaluación, que se denomina ontología de ajuste (ontology fit) y que fue propuesto por Brewster (Brewster et al., 2004), se emplea una ontología base, que es comparada con la ontología que se desea generar. La idea sobre la que se construye ese sistema es la de evaluar el corpus de los textos empleados en la creación de la ontología en relación con las instancias de dicha ontología, construyendo un clúster que permita valorar el nivel de consistencia semántica alcanzado.

Los resultados deberían revelar si la metodología empleada para el diseño de la ontología ha sido la correcta, o si el corpus con el que se ha trabajo cumple con los estándares correctos, pero no se centra demasiado en valorar el nivel de recuperación de información obtenido, o las inconsistencias semánticas (taxonomías mal formadas, errores en la ubicación de las clases, axiomas mal declarados, descripciones simplistas de las clases por el uso incorrecto de anotaciones, etc.). Un ejemplo claro de dificultad a la hora de aplicar esta técnica se produjo en el análisis de Wordnet (http://wordnet.princeton.edu/), donde existen instancias declaradas en diferentes partes de la ontología con evidente relación semántica entre ellas, pero sin que dicha relación esté implementada de forma alguna Fellbaum (Fellbaum, 1998). Según este sistema de evaluación, esa ontología tenía problemas metodológicos de construcción debido a ese tipo de relaciones, ya que era detectado como un problema estructural, cuando precisamente era uno de los fuertes de esta herramienta terminológica.

### 7.4.2.- Sistemas centrados en el costo

A nadie se le escapa que el proceso de construcción de una ontología, independientemente del modelo escogido (desde cero, con herramientas para la obtención automática, o semi-automática desde información textual, desde modelos entidad/relación, aprendizaje, etc.), es posiblemente el más costoso, desde todos los puntos de vista, dentro de la representación del conocimiento. Tanto es así, que existen numerosos sistemas que hacen especial hincapié en este apartado cuando se evalúa una ontología. En ocasiones, demasiado.

Teniendo en cuenta que existe una estrecha relación entre la ingeniería de software y la construcción de ontologías (dependiendo del modelo escogido para la creación se pueden llegar a solapar diversas etapas), no son pocos los autores que han defendido la idea de adoptar sistemas híbridos. Así por ejemplo, Ontocom (Paslaru Bontas and Mochol, 2005) emplea el sistema Cocomo (Constructive Cost Model) como base para construir un mecanismo de evaluación de ontologías. Emplea para ello la unión de tres metodologías propias de la ingeniería de procesos, como son top-down, parametric y expert-based Boehm (Boehm, 1981) para analizar el costo de construcción, mantenimiento y reutilización de la ontología teniendo en cuenta valores tales como el número de personas involucradas en la construcción, cantidad de horas dedicadas, tamaño de la ontología, etc.

El principal problema que presentan este tipo de sistemas se localiza en la total ausencia de análisis de la estructura taxonómica creada, la semántica aportada, las relaciones entre las propiedades, los axiomas, etc.

### 7.4.3.- Basados en métricas

Aquí recopilamos aquellos sistemas que utilizan diferentes cálculos, generalmente con base matemática, para valorar diversos aspectos de la ontología. Éste es el caso, por ejemplo, del modelo propuesto por Yao (Yao et al., 2005), donde se emplean un determinado número de métricas de cohesión específicamente creadas para su uso en ontologías. Están basadas en

diferentes teorías matemáticas y analizan, fundamentalmente, la profundidad de la descripción en función del número de clases principales, subclases, notaciones formales, profundidad media del árbol taxonómico, etc. Para llevar a cabo su propuesta evaluó varias ontologías con licencia creative commons por medio de un parser xml denominado OMP (Ontology Metrics Parser). Para la fase final del experimento se empleó a 18 evaluadores con dilatada experiencia en el proceso de creación de ontologías, no así de la materia que era representada en dichas herramientas.

Otros sistemas de evaluación también emplean metrías diferentes, y además con objetivos distintos. Así, por ejemplo, OntoMetric (Lozano-Tello and Gómez-Pérez, 2004), se centra más en analizar la posible reutilización de la ontología que la calidad de la misma. Para ello utiliza la jerarquía de procesos analíticos (Analytic Hierachi Process, AHP) como criterio para adoptar procesos de negocio a los ontológicos, analizando el contenido representado en la ontología, el lenguaje con que la ontología es implementada, la metodología con la que se llevó a cabo la construcción de la ontología, el software empleado y el costo de uso del sistema.

En total, este sistema analiza 160 características diferentes, convirtiéndose en una de las metodologías más exhaustivas. Quizá el principal problema que tenga es que, en ocasiones, está demasiado centrada en la reutilización de la ontología, y no en la evaluación como herramienta para la recuperación de información. En ese sentido, cabe destacar el trabajo de Sabou (Sabou et al., 2006) que emplea un conjunto de variables muy parecidas a las de OntoMetric, pero orientadas hacia ese aspecto. Destaca el análisis que realizó sobre el conocido motor de búsqueda Swoogle (http://swoogle.umbc.edu/).

### 7.4.4.- Sistemas basados en el análisis de jerarquías

Mediante este análisis, basado en la comparación, se facilita el estudio de la estructura conceptual de una ontología concreta para observar los posibles errores de exhaustividad, junto a otro tipo de restricciones. Emplea como

medida de evaluación el cálculo de distancias y la similaridad entre clases de las ontologías analizadas.

Es un mecanismo parecido al empleado en OntoClean (Guarino and Welty, 2002) que, además de establecer criterios de selección de ontologías a partir de una taxonomía base, analiza la posible existencia de problemas de inconsistencia semántica y/o estructural. Se centra especialmente en las propiedades definidas en la ontología, así como en su significado.

### 7.4.5.- Otros sistemas de evaluación

En alguna de las metodologías comentadas anteriormente, se hablaba de la existencia de observadores que, generalmente al final del proceso, se encargaban de aplicar los criterios de calidad. Existen otras opciones, como por ejemplo, que sobre estos observadores recaiga un peso mayor. Aunque no suele ser muy habitual, por el elevado nivel de subjetividad que implica, no es menos cierto que existen experiencias basadas en esta idea.

En el trabajo de Fernández-Breis (Fernández-Breis et al., 2009) se utiliza un sistema basado en la ISO 9126 que, aunque está pensada para la evaluación de la calidad del software, el autor logra adaptar al mundo de las ontologías, reconvirtiendo los 10 criterios generales de esta ISO en aspectos específicos del análisis de aspectos internos, externos y de uso de las ontologías. Una vez confeccionada la plantilla, 8 alumnos de doctorado realizan la evaluación de dos versiones diferentes de una ontología especializada en biología.

Resulta evidente que en este tipo de métodos, el peso que debe soportar la evaluación se encuentra sobre un terreno excesivamente subjetivo. Además de mencionar que criterios relacionados con la calidad y la cantidad de clases y sus relaciones no son analizados con la profundidad que se requiere. Algo que, por otra parte, no sucede con el sistema propuesto por Gangemi (Gangemi et al., 2005), quien emplea una meta-ontología (se trata de una propuesta meta-teórica sobre cómo se debe realizar la evaluación tanto de la ontología como de las anotaciones que incluya), denominada $O^2$ y una ontología específica

para la evaluación y validación, llamada oQual, que es la que incluye los parámetros a aplicar.

Hemos dejado para el final los métodos experimentales llevados a cabo por Evermann (Evermann and Fang, 2010), quien evalúa la ontología desde un paradigma cognitivo, muy apegado a los principios del desarrollo de aplicaciones.

Independientemente del esquema escogido para evaluar la ontología, lo normal es que se presente al final un listado de errores, que pueden ser achacados a su creación, implementación en el sistema de información, disonancia entre el objetivo y lo realizado, etc. A continuación pasaremos a comentar brevemente los resultados más habituales en este capítulo.

### 7.4.6.- Errores Frecuentes en el diseño de Ontologías

Lo normal es que la construcción de la ontología lleve, tarde o temprano, a algún tipo de error. La importancia de éste determinará al final el nivel de operatividad alcanzado con la misma. La tipología de errores en el diseño de ontologías va desde lo fallos en la taxonomía y diseño, hasta equivocaciones en la semántica conceptual, en la postura estructural y en los axiomas que se declaran, cuestiones que son causantes de bajos resultados cuando se realizan operaciones en los sistemas.

Son múltiples los errores que comprometen el desarrollo exitoso de una ontología. En el trabajo de Fahad (Fahad and Abdul Qadir, 2008) se describe un sinnúmero de anomalías en el diseño ontológico. Dichos errores de diseño y concepción pueden clasificarse como simples errores de construcción, pues evidencian problemas en el diseño de las instancias y en la construcción de las clases del sistema ontológico. Otros autores, Staab, por ejemplo (Staab, 2004) analizan los problemas de construcción de las ontologías y los pone en relación con los resultados que acarrean estos fallos. Recogemos a continuación la relación de errores más comunes, a partir de lo expuesto por Fahad (Fahad and Abdul Qadir, 2008), Staab(Staab, 2004), Ning (Ning and Shihan, 2006b)

Maedche (Maedche et al., 2003, Maedche and Staab, 2002b) y nuestra propia experiencia.

- Errores de ubicación: Se producen cuando no se tienen en cuenta la disposición correcta de las clases y se solapan estas, principalmente por la existencia de cruces entre nodos y estructuras conceptuales Ning (Ning and Shihan, 2006a). Suele suceder cuando una misma clase es definida como una subclase y como una superclase al mismo tiempo en diferentes niveles de jerarquía de la ontología. Este tipo de errores se producen cuando, a la hora de construir las relaciones semánticas, no se realizan análisis lingüísticos y/o terminológicos de los conceptos, provocando una estructuración del conocimiento débil o poco coherente.
- Errores de distribución. Suele aparece cuando, a la hora de estructurar el conocimiento, los procesos creados se realizan desde una base netamente clasificatoria. Es en ese momento cuando se puede producir una excesiva dependencia (genérico/específico o “tipo de” y “subtipo de”) entre las clases y las subclases (Ning and Shihan, 2006a). Como resultado de este error aparecen clases que están mutuamente atadas a subclases disjuntas(Qadir and Noshairwan, 2007b). Otra inconsistencia de este problema está en la fragmentación de una clase en muchas subclases, sin tener en cuenta que las instancias de las subclases no necesariamente tienen que pertenecer a las subclases que se han declarado.
- Errores de inconsistencia semántica. Se produce al desarrollar una jerarquía de nodos para un concepto erróneo. Es decir, cuando en la jerarquía aparecen conceptos que no pertenecen a la clase principal. La razón principal de este error es el poco conocimiento semántico y terminológico del área del conocimiento con la que se está trabajando.
- Clases y clasificaciones incompletas. En ocasiones se presta muy poca atención a elementos muy importantes en la descripción, en la anotación y en la conceptualización de la ontología, cuestión que favorece la ambigüedad y entorpece la construcción de herramientas de razonamiento. Esto repercute en clases o clasificaciones poco

documentadas. A este tipo de errores también pertenece el siguiente grupo.

- Omisión de conocimiento disjunto. Errores que ocurren cuando existe un gran número de clases y subclases, y en su creación se obvia la inclusión de axiomas, que tienen determinado grado de disjunción. Según Qadir (Qadir and Noshairwan, 2007a) estos errores aparecen en la fase de diseño debido a la falta de conexión entre el desarrollador y los usuarios, lo que hace que se obtengan resultados catastróficos.
- Omisión de conocimiento por falta de exhaustividad. Se produce cuando los constructores de la ontología no crean restricciones de integridad, declarando de forma arbitraria particiones y subclases de un mismo concepto. Con esto se deja en el aire la capacidad de exhaustividad del sistema, conduciendo a la redundancia.
- Errores de redundancia. La repetición de conceptos, lo que evidencia la falta de un plan en la confección de las taxonomías de base del sistema, como apunta Fahad (Fahad and Abdul Qadir, 2008).
- Errores en la Poca especificación o delimitación de las propiedades de los componentes del sistema. Esto produce que el razonamiento se vea poco desarrollado, ya que se da una misma definición a todo, desde el punto de vista formal y conceptual.
- Errores por falta de exhaustividad en la declaración de etiquetas. Este error está muy asociado a la construcción de inferencias en las ontologías que generan su estructura usando OWL. Las etiquetas que deben ser desarrolladas o declaradas deben tener suficientes elementos y atributos para detallar correctamente un dominio y hacer que los conceptos puedan estar asociados a determinadas propiedades. En algunos sistemas ontológicos se pasa por alto esta propiedad, lo que en hace que los modelos de razonamiento se vean desarticulados. Los problemas esenciales que se describen por este error son del tipo IFPO (Omisión de las Propiedades) es decir, que si no se declaran estos valores el sistema sólo identifica un único valor para un concepto y una sola clave en la base de datos.

- Errores por no describir correctamente el conocimiento. Lo que acarrea que no se reconozca exactamente qué elemento se está declarando en cada concepto, ya que no hay definiciones exactas para subclases y sus relaciones con los conceptos (Maedche and Staab, 2002a). Esto obliga a una necesaria descripción de reglas que proporcionen criterios esenciales para la interacción con nuevos conceptos Kuhn (Kuhn, 1979b) que se creen en las subclases y sus relaciones, es decir obliga a generar modelos de axiomas, complementaciones y restricciones para el trabajo con la ontología.
- Errores de redundancia en las extensiones. Aparece cuando se describe un concepto como disjunto con otro concepto de la misma jerarquía (Kuhn, 1979a, Maedche and Staab, 2002a).

### 7.5.- Evaluación de Onto-Satcol

Onto-Satcol es una ontología desarrollada para formalizar parte del conocimiento asociado al dominio de la Ingeniería de Puertos y Costas. Su nacimiento viene provocado por dos motivos: como herramienta para un sistema de desambiguación de textos científicos, y para dar respuesta a la necesidad de paliar el desigual aumento de información en lo que se refiere a la existencia de herramientas terminológicas que faciliten la recuperación de información (taxonomías, tesauros y ontologías) dentro de esta área. Con estas características, es Ontosatcol una ontología única para el dominio de la Ingeniería de Puertos y Costas, cuestión por la cual fue necesario desarrollar un modelo de evaluación específico para realizar su valoración antes que fuera parte del sistema de PUERTOTEX. Los criterios de evaluación que se presentan aquí están concebidos a partir del estudio teórico de los postulados de Recuperación de la Información y el desarrollo de Ontologías. Con el fin de evaluar Onto-Satcol se utilizaron diversas métricas, sobre todo aquellas que facilitaran la evaluación de su diseño y aplicación en los procesos extractivos y semánticos.

### 7.5.1.- Onto-Satcol: elementos constitutivos y características

Se trata de una ontología de elevado nivel de complejidad, tanto por la temática como por su desarrollo. Posee 6.284 conceptos en español y 23.546 instancias, formalizadas en OWL, al igual que los 89 axiomas que facilitan la estructuración del conocimiento. Esta ontología posee además imágenes que están asociadas a los conceptos, así como una estructura conceptual que deriva relaciones paradigmáticas de elevado nivel de complejidad al igual que Wordned y EuroWordned.

Los términos de la ontología están asociados a subclases que enuncian su estructura semántica (término, sustantivo, adjetivo, homónimo, merónimo, hipónimo, hiperónimo, etc.). Cada término definido como clase principal tiene asociado una propiedad que se denomina, translate, es decir, un conjunto de equivalencias para ese término en inglés y alemán. Se prefirió este camino al uso de las anotaciones para enriquecer las futuras búsquedas en otros idiomas. A la vez, cada término tiene asociada una imagen, almacenada en el sistema junto a su anotación, definida por las siguientes etiquetas:

- **Connotacional:** información sobre lo que connota el objeto.
- **Denotacional:** elementos que facilitan declarar aquellas características que se declaran en la imagen: dimensiones, estado físico, técnica de reproducción y obtención. etc.
- **Keyword**: palabras clave.
- **URL:** ubicación donde se encuentra ese archivo
- **Autor:** responsable del documento y de su información.
- **Datos onomásticos:** personalidades, etc. Incluidas en la imagen.
- **Datos topográficos:** lugar físico de donde se ha tomado la imagen.

El sistema ontológico trabaja con textos estructurados en RDF Schema, y trabaja con etiquetas que están en consonancia con la ontología, con lo que se facilita que un sistema automático (generalmente llevado a cabo por agentes)

realice acciones concretas en el documento (localizar nuevos documentos, conectar enlaces a clases, etc.) (Ver figura 138).

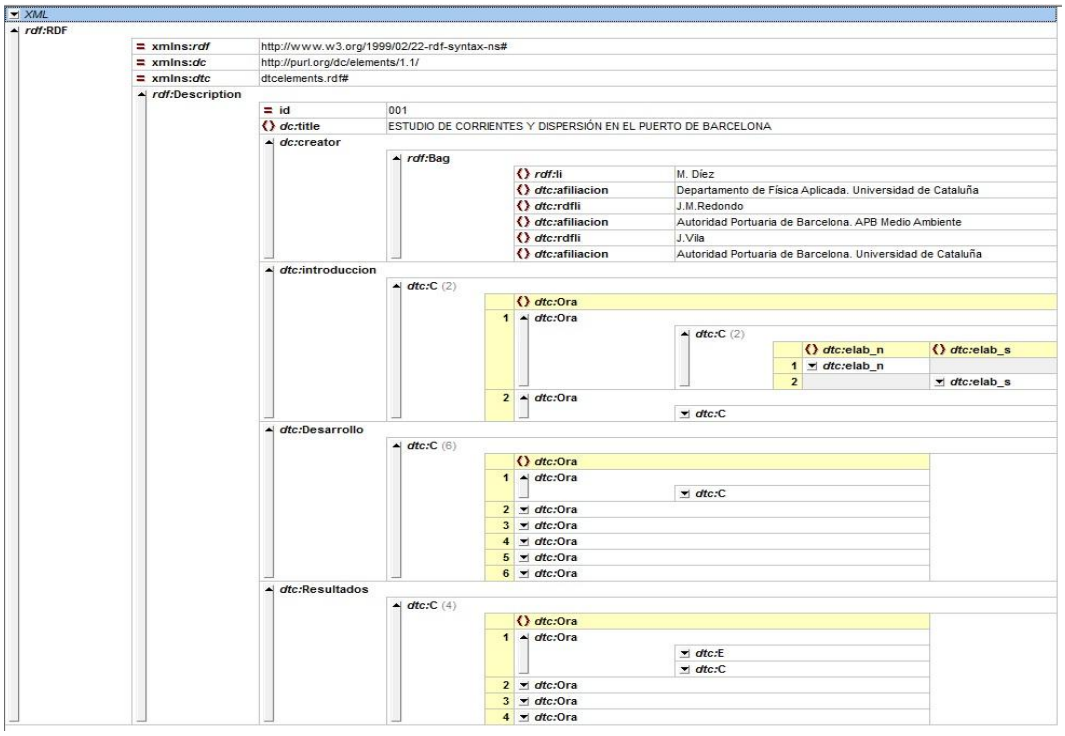



Figura 138. Estructura sintáctica de los documentos de la ontología en RDF con XML Spy

La ontología cuenta también con un modelo de desambiguación de textos asociado a partir de las palabras que se marcan en el texto. Dicho modelo, empleado también para la recuperación de la información, se implementó con un sistema de anotación constituido por un conjunto de metadatos Dublin Core junto a otro conjunto de etiquetas ad-hoc que tienen como finalidad el facilitar los procesos de recuperación, extracción, desambiguación y equivalencia semántica de los textos.

Este elevado nivel de complejidad obliga a que los procesos de evaluación de la aplicación sean desarrollados de forma diferente a los propuestos hasta la fecha, y ya comentados en el apartado 2. Se ha desarrollado un conjunto de medidas o métricas construidas a partir de las necesidades de los disímiles procesos que asume esta gran ontología. Los indicadores de análisis de la

ontología se han desarrollado a partir de las mismas métricas que se conocen, pero se les ha dado una connotación especial para adaptarlas a la aplicación que se desarrolla.

#### 7.5.1.1.- Variables para la evaluación: análisis

Los indicadores para la evaluación de esta ontología están pensados para evaluar los objetivos de la misma dentro de un sistema de extracción y desambiguación de textos, con el fin de introducir mejores para futuras aplicaciones y sistemas. Para evaluar la ontología se han declarado indicadores léxicos, de recuperación de la información, de la estructura sintáctica y una fase de constatación.

#### 7.5.1.2.- Variables léxicas

Miden la calidad del léxico en lo referente a la cobertura temática, exhaustividad de la anotación, redundancias semánticas, taxonomías ambiguas, capacidad de desambiguación y capacidad de traducción. A continuación se declaran las ecuaciones y la explicación de cada indicador:

**Cobertura temática:** una ontología tiene cobertura temática suficiente si formaliza la mayoría de los conceptos de la esfera del conociendo que representa. En el caso de los sistemas que simbolizan una materia o conocimiento nuevo habría que verificar cuál es el nivel de alcance de la ontología en cuanto a conceptos dentro de esta ciencia nueva. Se desarrolla con la fórmula Ct= CCO / CCM

Donde CCM representa las temáticas y CCO es la cantidad de conceptos que posee la ontología.

**Exhaustividad de la anotación:** valora el nivel de las anotaciones y su valor semántico en el momento de representar la información. La fórmula emplea es: EA = CNC / CN

donde EA es la exhaustividad de la anotación, CNC es la cantidad de notas correctas y CN es la cantidad de notas de la ontología.

**Redundancias semánticas**: la fórmula usada es: RSE = CCA / CCO

donde RSE son las redundancia semánticas, CCA la cantidad de conceptos ambiguos y CCO representa a la cantidad de conceptos de la ontología.

**Capacidad de desambiguación:** se calcula con la fórmula CDA = NTDA/NTD

donde CA es capacidad de desambiguación, NTDA es número de etiquetas desambiguados en un documentos y NTD responde al número de etiquetas marcadas en el documento. Para su aplicación se calcula la sumatoria de cada resultado de cada etiqueta desambiguada, entre las etiquetas de cada uno de los documentos de la colección de estudio declarado.

**Capacidad de traducción:** Es la capacidad que tiene el sistema de realizar equivalencia en el texto con aquellos términos que poseen una etiqueta como marca, se denota aquí como CT= CTT/CTE

donde CTT es la cantidad de términos traducidos dividido entre la cantidad de términos equivalentes (CTE).

#### 7.5.2.1.- Variables de recuperación de información

Estos indicadores miden la efectividad de la ontología para recuperar información relevante para el usuario, a partir de las medidas clásicas de recuperación de la información: precisión y exhaustividad. Dado que estas medidas están pensadas para la evaluar la recuperación de información, hemos desarrollamos una variación con el fin de que se pueda adaptar específicamente a la ontología que estamos analizando.

**Precisión en la recuperación de documentos:** esta medida estudia la capacidad del sistema para recuperar documentos relevantes a una petición de

los usuarios. La fórmula es PrecD = CDR / CDO, donde CDR es cantidad de documentos recuperados y CDO cantidad de documentos en la ontología.

**Precisión en la recuperación de documentos de tipo2**: es la capacidad de la ontología de devolver imágenes relativas a algún elemento lingüístico, y se calcula con PrecD2 = CIR/CIO, donde CIR es la cantidad de imágenes recuperadas y CIO es la cantidad de imágenes que han sido descritas en la ontología.

**Exhaustividad:** Es la proporción de material relevante recuperado, del total de objetos relevantes de la ontología, y se calcula con la fórmula ExausD = DR / CDOst, donde DR es la cantidad de objetos relevantes a la solicitud y CDOst, es la cantidad de objetos que existe sobre ese tema en la ontología.

#### 7.5.1.3.- Variables para la evaluación de la estructura sintáctica

Para evaluar la ontología a nivel sintáctico utilizamos Protex, una herramienta desarrollada específicamente para evaluar este sistema de información. Protex de Domínguez (Domíguez, 2010) está desarrollado a partir de un algoritmo que genera un parser y un analizador sintáctico escrito en Phyton, permitiendo la visualización de los elementos de una taxonomía al igual que las herramientas Case (Computer Aided Software Engineering) desarrolladas a tales efectos. El objetivo de esta fase del análisis es la localización de incumplimientos en los estándares ontológicos como OWL (Ontology Web Lenguaje), RDF (Resoursce Description Framework) y DAML (Darpa Agent Markup Lenguaje). Protex fue creado con la finalidad de permitir localizar los siguientes errores:

- Identificación de solapamiento en el desarrollo de clases
- Localización de conceptos vacíos
- Omisión de conocimiento disjunto
- Omisión de conocimiento por falta de exhaustividad
- Conceptos mal formulados

## 7.6. – Adecuación de los requerimientos

Como en todo sistema ontológico que se evalúa, siempre se hace imprescindible valorar los resultados del proceso. Para ello es obligatorio decidir si la ontología cumple claramente con los estándares internacionales y si los resultados que arroja la investigación empírica tienen la validez necesaria para ser tomados en cuenta, de lo contrario hará que comenzar desde el inicio para valorar las deficiencias.

### 7.6.1.- Experimento

Para llevar a cabo la evaluación se han escogido a 5 expertos en el tema de la Ingeniería de Puertos y Costas. En otros métodos, como los comentados anteriormente, se suele elegir principalmente a expertos en ontologías antes que a conocedores del área de conocimiento. Entendemos que esa opción aporta una visión excesivamente sesgada de la evaluación, muy centrada en la ontología y poco en el grado de corrección de la organización del conocimiento plasmada, y por eso hemos optado por esta otra. Ellos serán los que analizarán los documentos y la estructura semántica de la ontología, y verificarán la precisión y el exhaustividad de cada una de las preguntas empleando los indicadores anteriormente descritos. Para llevar a cabo su trabajo emplearon Protex ya que, por un lado permite la navegación dentro de la ontología, y por otro tiene implementados gran parte de los indicadores, por lo que puede generar a partir de ellos los informes para proceder a la evaluación.

## *7.7.- Resultados de la Evaluación de Ontosatcol*

### 7.7.1.- Indicadores léxicos

Se valoran a partir de los datos aportados por los observadores (Anexo 2), analizando el nivel de cobertura temática alcanzada en los nodos principales de la ontología. Cada experto anota el número de nodos temáticos que cree se corresponden con el dominio que representa. Los resultados han sido calculados de forma individual para cada uno de los procesos, añadiendo una

variación en la fórmula básica para el cálculo final. El cálculo de la cobertura temática (Ct= CCO/CCM) será la media de resultados aportados por cada observador (Ct1=CCOb1/CCM). En nuestro caso, el resultado final ha sido 0,7 lo que indica que solo un 70 % de los conceptos relativos a ese dominio están representados conceptualmente en esa ontología.

Para el cálculo de la exhaustividad en la anotación los expertos dividieron la ontología en 6 secciones, y tenían como objetivo valorar todas las notas que no estaban correctas. De las 23.546 instancias sólo 376 estaban correctas, lo que otorga un valor de 0,16. Esto indica que las anotaciones no son eficientes, lo que puede impedir una buena recuperación de la información y provocar una baja relevancia.

Para analizar las redundancias de significado se utilizó la estructuración Sense de la herramienta PROTEX, la cual le brindo a los expertos la posibilidad de visualizar los nodos para determinar, mediante marcado, aquellas redundancias semánticas. El índice resultante fue de 0,15, lo que demuestra que el nivel de representación en una buena cantidad de términos es relativamente bajo. Ese importante nivel de ambigüedad conceptual es muy importante, posiblemente debido a que la extensión de la ontología, lo que nos indica que es necesario hacer cambio en la estructura de la misma con el fin de tener mejores resultados. No obstante, dependiendo del nivel de profundidad de los nodos se observaba una varianza evidente en cuando a la ambigüedad se refiere. Esto podía ser debido a que algunas subáreas de conocimiento estaban mal representadas, pero no todas. Por ese motivo realizamos un test de desambiguación.

Con el fin de conocer si la ontología se podría emplear como herramienta para desambiguar textos especializados se han seleccionado 50 documentos de ese tipo debidamente etiquetados con RDF. En cada uno de ellos se han marcado palabras claves, así como verbos, sustantivos, adjetivos y frases semánticas. Los textos fueron sometidos a desambiguación, activando la opción desambiguar que posee la herramienta Protex. De las 890 palabras marcadas

en cada texto se desambiguaron 789, lo que indica que el sistema es bastante fiable en cuanto a desambiguación y trasformación del corpus, es decir 0,88 es el nivel de desambiguación, un valor bastante elevado para un sistema con tan elevada complejidad.

Para el estudio de la capacidad de traducción se analizaron 50 textos, en los cuales aparecían 150 marcas de equivalencia, que permitieron la traducción de 118 términos, lo que evidencia que se consigue un índice elevado de traducción 0,78 en función de la muestra (Tabla 70).

| Indicadores | Valor |
|---|---|
| Cubrimiento Temático | 0,7 |
| Exhaustividad de la anotación | 0,16 |
| Redundancias semánticas | 0,15 |
| Capacidad de desambiguación | 0,88 |
| Capacidad de traducción | 0,78 |

Tabla 70. Indicadores de Evaluación Léxica

### 7.7.2.- Precisión en la recuperación de documentos

Para analizar este indicador se les pidió a los expertos que realizaran 15 búsquedas en el sistema y que midieran la relevancia de los documentos recuperados. Con el fin de dificultar más el proceso, y aumentar el riesgo de ruido, las consultas estaban formadas por un único término. El resultado fue que de los 2.890 documentos de media recuperados, 2.345 eran relevantes, evidenciando un elevado nivel de relevancia, un valor medio de 0.81. En cuanto a la recuperación de documentos de tipo dos, es decir material no textual, se recuperó por tipo de documentos todas las fotos del sistema para las 15 peticiones, aportando un valor de 1. Esto indica que el sistema es confiable en lo que respecta a precisión cuando recupera un material no textual.

### 7.7.3.- Indicadores de evaluación de la estructura sintáctica

La herramienta empleada para estos cálculos, Protex, descubrió la existencia de 26 clases que poseen un nombre diferente, pero un contenido (expresado

en la anotación) similar, por tanto se solapan. Se han visualizado 69 conceptos vacíos, es decir con errores de significado, eso crea problemas semánticos y hace más compleja la taxonomía. En la ontología se han encontrado 89 errores al tratar conocimiento disjunto, es decir hay 67 conceptos macro, que poseen igual número de sub-conceptos con igual estructura jerárquica. De igual forma aparecen 67 problemas por conceptos mal descritos por falta de calidad en la exhaustividad en la descripción de los documentos (figura 139).

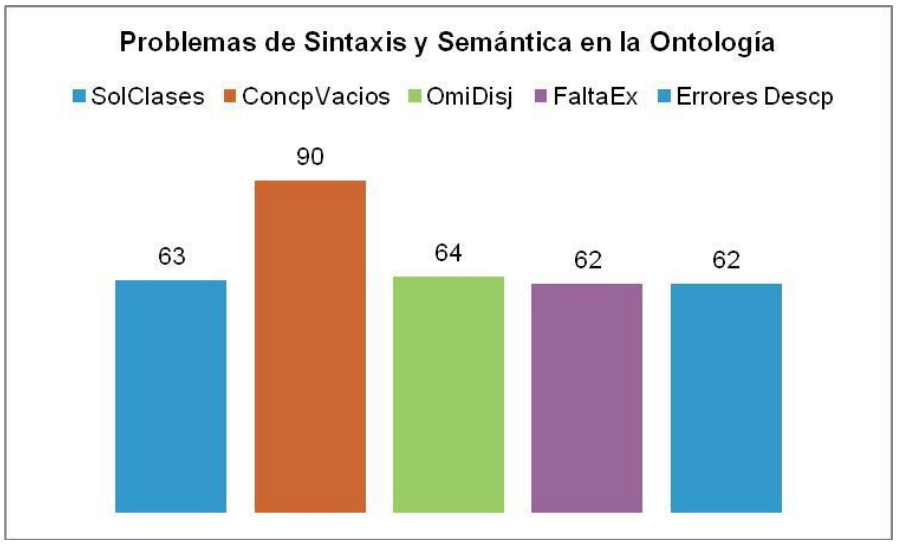



Figura 139. Problemas de sintaxis y semántica

### 7.4.- Adecuación de los requerimientos

Este experimento no puede terminar hasta que no exista una verificación de la exactitud de los datos y la calidad de los mismos. Para ello se decidió formular algunas preguntas que facilitarán la detección de los conceptos, la traducción y la desambiguación (Anexo 8). Estas preguntas fueron las siguientes:

1. ¿Cuáles son los efectos que produce la marea en la Costa de Cantabria?
2. ¿Cuáles son las centrales azucareras más contaminantes de Cuba?
3. ¿Con qué sustancia química se hace la espectrografía de gases?
4. ¿Cuáles son las zonas geográficas de menor desarrollo portuario?
5. ¿Cuál es la traducción del término “rompiente de derrame” en alemán?

6. ¿Cuál es su contexto de aplicación?
7. ¿Cuántas imágenes de derrame de petróleo existen en la ontología?

Los resultados de estas preguntas demostraron valores muy similares a los observados durante los experimentos, con lo que queda demostrada la veracidad de la información proporcionada por el experimento (Ver Tabla 71). Las respuestas a estas interrogantes fueron contestadas por la ontología con los siguientes niveles:

| Preguntas | Exhaustividad | Precisión |
|---|---|---|
| Pregunta 1 | 0,5 | 0,832 |
| Pregunta 2 | 0,3 | 0,901 |
| Pregunta 3 | 0,5 | 0,700 |
| Pregunta 4 | 0,7 | 0,892 |
| Pregunta 5 | 0,1 | 0,823 |
| Pregunta 6 | 0,1 | 0,992 |
| Pregunta 7 | 0,2 | 0,657 |

Tabla. 71. Evaluación de la Exhaustividad y la Precisión

### *7.8.1.- Evaluación de Resúmenes mediante Rouge*

Los resultados que se presentan aquí han sido previamente evaluados por los resumidores humanos. Los textos han sido incluidos en nuestra herramienta Protex (Domínguez, 2010), diseñada para el etiquetaje semiautomático de los corpus y sus oraciones. Esta evaluación se hace con el objeto de analizar resúmenes desarrollados bajo diversos modelos de tratamiento textual, como los de microsft Word, baseline (Anexo 32), puertot_1ª (Anexo 28), 1b (Anexo 29) y puertot_2ª (Anexo 30), 2b (Anexo 31).

La métrica más reconocida para evaluar en el Resumen o Extracto es Rouge[53] (Lin and Hovy, 1998). Atendiendo a que en esta investigación se utiliza una técnica para desarrollar resúmenes automáticos, es imprescindible actuar con Rouge.

[53] Rouge es una herramienta de evaluación de textos automáticos generada por la DUC (Document U Conference

#### 7.8.1.1.- Resúmenes Candidatos

Cuando se trabaja con Rouge se necesita tener resúmenes candidatos y resúmenes de comparación. Para lograr tales requerimientos en la evaluación, se seleccionaron 10 resúmenes de baseline y 10 resúmenes de obtenidos a partir de Puertoterm (Anexo 28).

Los sumarios objeto de evaluación son 10, desarrollados con el fin de actuar como baseline (Anexo 32). En cada texto resumido se han seleccionado de forma aleatoria un número de oraciones que representan la estructura retórica del texto primario. Estas oraciones se seleccionan con un programa de selección de números aleatorios.

Al igual que en D'Cunha (D'Cunha, 2006) se utilizaron resúmenes generados a partir de Microsoft Word (Anexo 33) para realizar resúmenes de 25 oraciones, que sirvan de candidatas para la evaluación.

### 7.8.2.- Resúmenes de Referencia

Es vital cuando se trabaja con Rouge, poseer resúmenes que sirvan de patrón o referencia para ser comparados con los resúmenes base o candidatos. En este experimento las comparaciones se realizan atendiendo a los siguientes tipos de resumen:

- Resúmenes por extracción generados por el sistema.
- Resúmenes por abstracción generados por el sistema.
- Resúmenes por extracción realizados por 10 especialistas de la Comunidad Ciencias Biológicas.
- Resúmenes por abstracción realizados por 10 especialistas de la Comunidad Ciencias Biológicas.

La mayoría de los autores prefieren evaluar con Rouge utilizando como materia de análisis solo los resúmenes de los autores. La experiencia internacional en la evaluación con esta métrica aclara que no es fiable evaluar un resumen

automático usando un solo tipo de resumen obtenido bajo un solo modelo, por tanto en esta evaluación se introducen otros resúmenes para ser analizados.

La metodología para obtener estos resúmenes a partir de los procesos de extracción y abstracción desarrollados en la investigación se encuentra declarada en el apartado (5.5), y se especifica en este acápite por ser el que concierne a la valoración. Se debe aclarar que los artículos que han de evaluarse no poseen todas las mismas extensiones, son originales y su fecha de publicación abarca los años 2000-2007.

Los corpus de contraste de la investigación están formados por los siguientes grupos de textos:

I) **Subcorpus Manual de Contraste**

- Artículos originales del tema Ingeniería de Puertos y Costas (10).
- Resúmenes indicativos realizados por autores especializados en el tema (10).
- Resúmenes construidos bajo la técnica de extracción de oraciones (10).

II) **Colección de Entrenamiento para Validación Final**

- artículos científicos especializados en Ingeniería de Puertos y Costas en español.
- resúmenes indicativos realizados por los autores sobre los referidos artículos.

#### 7.8.2.1.- Acopio de Elementos Textuales

Este acápite explica cómo se reúne el texto que ha de ser utilizado para la evaluación y como se procesa dicho texto, antes de su valoración. Se indicará a los especialistas realicen las siguientes acciones:

1. Delimitar las oraciones que posean mayor relevancia en los textos científicos que se les entregan. El tope máximo de oraciones de análisis es 25, esto permitirá tener un resumen por extracción.

2. Teniendo en cuenta las oraciones seleccionadas en el paso anterior, redactar un resumen indicativo teniendo en cuenta las normas (UNE, 1990)De esta forma se obtiene un resumen por abstracción.

#### 7.8.2.2.- Organización de los Datos Registrados

Los datos obtenidos en la fase anterior deben ser organizados con vistas a su posible evaluación con Rouge y sus variantes. Los pasos de Recogida son los siguientes:

1. **Resumen por Abstracción**

    - Extraer las oraciones que se han delimitado en el texto.
    - Se colocan siguiendo el orden del texto fuente.
    - Archivar en una carpeta los diez resúmenes desarrollados por cada médico por separado.

2. **Resumen por Abstracción**

    - Teclear en Microsoft Word los resúmenes confeccionados por los expertos.
    - Crear una carpeta con los resúmenes hechos por cada experto por separado.

Estos corpus se incorporan a la base de datos Textos, y luego son tratados y marcados semi-automáticamente y almacenados en la referida base de datos.

7.8.3.- Aplicación de Rouge

Teniendo bien delimitados todos los textos candidatos y de referencia estamos en condiciones de aplicar ROUGE. Rouge está diseñado para evaluar resúmenes formalizados en idioma inglés por tanto en esta valoración de calidad fue necesario realizar las mismas adaptaciones que aparecen generadas en la tesis de D'Cunha y Lin y Hovy (D'Cunha, 2006, Lin and Hovy, 1998).

- Se sustituye la lista de palabras vacías en inglés por la lista stop word elmination obtenida con Word Smith Toolls.
- Se suprime el stemmer de Porter, pues sus bondades están en función del inglés
- Se utilizaron resúmenes lematizados previamente utilizando el TreeTagger para el español.

### 7.8.3.1.- Rouge: Variantes

Rouge de Lin (Lin and Hovy, 1998), es una herramienta para la evaluación de la sumarización de documentos y se ha convertido en una de las pautas esenciales para determinar la calidad de los resúmenes realizados mediante método automáticos. Rouge posee varias métricas para realizar la valoración de los resúmenes, esto facilita la selección de registros adecuados para su adaptación a cada proyecto. Las métricas que evalúa Rouge son las siguientes:

- **Rouge N (N-Gram Co-Ocurrence Statistic):** Esta media estudia la cantidad de ngramas que ocurren en un resumen candidato y en un resumen de referencia. Las referencias obtenidas en el numerador de la ecuación indican las sumas de todas las referencias declaradas en el resumen base, sin embargo, los resultados que se obtienen el denominador indican que se asumen más referencias al texto obtenido mediante métodos estadísticos. La ecuación declara que Rouge N es el cociente de la suma de las referencias del resumen base entre las referencias de un resumen obtenido mediante medidas automáticas. En ella $Count_{match}\ (gram_n)$ es el número máximo de referencias recibidas por un resumen candidato y un resumen de referencia. Los n-gramas son medidas resultantes de n, que significa la cantidad de n-gramas que ocurren en ambos textos.

$$\frac{\sum_{S \in \{Referemce\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- **Rouge-L (Longest Common Subsequence):** A partir de dos secuencias X y Y se definen la mayor sub-secuencia para ambos (LCS). Estas operaciones han sido utilizadas por Saggion y Lapalme (Saggion and Lapalme, 2002) para comparar la similaridad de dos sumarios obtenidos mediante técnicas automáticas. La aplicación de (LCS) en la evaluación de la sumarización de textos ha permitido computar un texto como una secuencia de palabras y una (LCS) de base para calcular Rouge-L a partir del radio que se genera de las longitudes de las sub-secuencias de dos sumarios. LCS se calcula también usando un sumario de referencia y un sumario base. Las ventajas de LCS estriban en que no requiere del uso de constantes análisis ni observaciones para reflejar el orden de las palabras en las oraciones como n-gramas. Desde el punto de vista automático con LCS se logra incluir las subsecuencias comunes y los n-gramas, sin tener que definirlos.

- **ROUGE-W (Weighted Longest Common Subsequence):** Usando LCS clásico, dejan de analizarse las relaciones espaciales que se dan cuando existen subsecuencias alojadas dentro de otras subsecuencias. Es por ello que se introduce esta métrica llamada ROUGE-W, determinada a través de LCS con consecutivas observaciones. ROUGE-W puede ser implementado eficientemente si se usa programación dinámica Lin y Hovy (Lin and Hovy, 1998).
- **ROUGE-S (Skip-Bigram Co-occurrence Statistics):** Con Rouge-S se obtienen los bi-gramas que ocurren en dos textos, uno de referencia y otro candidato. Si se compara Rouge-S con Rouge LCS puede verse que LCS solo analiza las subsecuencias comunes, suprimiendo en su análisis expresiones que pueden favorecer combinaciones terminológicas como *en la*, etc., lo que limita el nivel máximo de distancia, algo que si se consigue con Rouge-S, a través del orden de

las palabras y de la forma skip-bigram. Rouge-S con el máximo de distancia de N es llamado ROUGE-SN. En la oración siguiente puede verse el cálculo de Rouge *"Los sistemas que analizan el sulfuro son muy costosos" C= (9,2)$^4$=11*.

- **ROUGE-SU: Extension of ROUGE-S:** Es una extensión de Rouge-S. Esta métrica está basada en las incapacidades operativas de Rouge-S, ya que no da crédito a aquellas oraciones donde no exista un par de palabras que co-ocurran en determinada referencia. Rouge-S adiciona los unigramas contándolos como una unidad de referencia.

### *7.8.3.2.- Resultados de la Aplicación de Rouge*

Como se ha visto desde esta sección de evaluación de resúmenes, el autor, intenta analizar la calidad del modelo que se propone en esta tesis mediante Rouge. Rouge en esta investigación se centra en la relación física de un texto básico y un texto de referencia desarrollado por un ser humano. Cuestiones como coherencia y cohesión van a ser evaluados en otra sección de este capítulo de evaluación. El autor utilizó todas las variantes de Rouge, para obtener una valoración completa de la calidad obtenida en los resúmenes (Ver Tabla 72- 76).

| Rouge-N | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6891 | 0.6734 | 0.6791 | 0.6732 | 0.6791 |
| Puertot_1b | 0.6781 | 0.6234 | 0.6451 | 0.6831 | 0.6321 |
| Puertot_2a | 0.6893 | 0.6129 | 0.6234 | 0.6612 | 0.6523 |
| Puertot_2b | 0.6321 | 0.6321 | 0.6247 | 0.6231 | 0.6643 |
| Baseline | 0.1345 | 0.1245 | 0.1891 | 0.1789 | 0.6745 |
| Word | 0.4567 | 0.4321 | 0.4831 | 0.4890 | 0.6267 |
| Rouge-N | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6589 | 0.6128 | 0.6238 | 0.6345 | 0.6538 |
| Puertot_1b | 0.6635 | 0.6238 | 0.6363 | 0.6136 | 0.6156 |
| Puertot_2a | 0.6123 | 0.6341 | 0.6721 | 0.6721 | 0.6825 |
| Puertot_2b | 0.6621 | 0.6458 | 0.6821 | 0.6239 | 0.6523 |
| Baseline | 0.1980 | 0.1213 | 0.1890 | 0.1992 | 0.1389 |
| Word | 0.4467 | 0.4125 | 0.4712 | 0.4529 | 0.4578 |

Tabla. 72 Rouge N gramas Clásico

| Rouge-2 | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6791 | 0.6714 | 0.6711 | 0.6722 | 0.6711 |
| Puertot_1b | 0.6718 | 0.6214 | 0.6421 | 0.6821 | 0.6316 |
| Puertot_2a | 0.6873 | 0.6119 | 0.6224 | 0.6622 | 0.6513 |
| Puertot_2b | 0.6311 | 0.6311 | 0.6227 | 0.6221 | 0.6633 |
| Baseline | 0.1335 | 0.1235 | 0.1881 | 0.1779 | 0.6735 |
| Word | 0.4557 | 0.4311 | 0.4841 | 0.4880 | 0.6257 |
| Rouge-2 | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6579 | 0.6118 | 0.6280 | 0.6335 | 0.6438 |
| Puertot_1b | 0.6625 | 0.6228 | 0.6353 | 0.6126 | 0.6256 |
| Puertot_2a | 0.6113 | 0.6331 | 0.6711 | 0.6721 | 0.6725 |
| Puertot_2b | 0.6611 | 0.6448 | 0.6811 | 0.6229 | 0.6513 |
| Baseline | 0.1970 | 0.1223 | 0.1890 | 0.1982 | 0.1379 |
| Word | 0.4457 | 0.4115 | 0.4702 | 0.4519 | 0.4568 |

Tabla. 73. Rouge-Bigrama Clásico

| Rouge/SU4 | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6701 | 0.6724 | 0.6731 | 0.6752 | 0.6731 |
| Puertot_1b | 0.6714 | 0.6224 | 0.6431 | 0.6801 | 0.6356 |
| Puertot_2a | 0.6893 | 0.6159 | 0.6244 | 0.6662 | 0.6573 |
| Puertot_2b | 0.6341 | 0.6311 | 0.6227 | 0.6221 | 0.6633 |
| Baseline | 0.1355 | 0.1235 | 0.1881 | 0.1779 | 0.6735 |
| Word | 0.4577 | 0.4311 | 0.4841 | 0.4880 | 0.6257 |
| Rouge /SU4 | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6599 | 0.6268 | 0.6580 | 0.6635 | 0.6538 |
| Puertot_1b | 0.6665 | 0.6528 | 0.6653 | 0.8126 | 0.6456 |
| Puertot_2a | 0.6133 | 0.6132 | 0.6511 | 0.6121 | 0.6625 |
| Puertot_2b | 0.6671 | 0.6648 | 0.6411 | 0.6829 | 0.6413 |
| Baseline | 0.1971 | 0.1423 | 0.1090 | 0.1182 | 0.1079 |
| Word | 0.4448 | 0.4515 | 0.4902 | 0.4219 | 0.4168 |

Tabla. 74. Rouge-Cuatri-grama Clásico

| Rouge-L | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6721 | 0.6824 | 0.6991 | 0.6952 | 0.6631 |
| Puertot_1b | 0.6614 | 0.6724 | 0.6641 | 0.6501 | 0.6556 |
| Puertot_2a | 0.6593 | 0.6359 | 0.6844 | 0.6862 | 0.6873 |
| Puertot_2b | 0.6741 | 0.6811 | 0.6927 | 0.6321 | 0.6733 |
| Baseline | 0.1355 | 0.1435 | 0.1781 | 0.1679 | 0.1935 |
| Word | 0.4777 | 0.4311 | 0.4991 | 0.4380 | 0.6557 |
| Rouge-L | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6999 | 0.6468 | 0.6980 | 0.6735 | 0.6738 |
| Puertot_1b | 0.6865 | 0.6728 | 0.6853 | 0.6826 | 0.6756 |
| Puertot_2a | 0.6933 | 0.6832 | 0.6511 | 0.6521 | 0.6825 |
| Puertot_2b | 0.6771 | 0.6948 | 0.6711 | 0.6629 | 0.6513 |
| Baseline | 0.1071 | 0.1523 | 0.1390 | 0.1682 | 0.1279 |
| Word | 0.4448 | 0.4715 | 0.4902 | 0.4219 | 0.4168 |
| **Rouge-W** | **Tex 1** | **Tex 2** | **Tex 3** | **Tex 4** | **Tex 5** |
| Puertot_1a | 0.6833 | 0.6924 | 0.6991 | 0.6692 | 0.6931 |
| Puertot_1b | 0.6945 | 0.6024 | 0.6641 | 0.6811 | 0.6856 |
| Puertot_2a | 0.6834 | 0.6659 | 0.6844 | 0.6972 | 0.6893 |
| Puertot_2b | 0.6941 | 0.6811 | 0.6927 | 0.6421 | 0.6773 |
| Baseline | 0.1855 | 0.1435 | 0.1781 | 0.1979 | 0.1265 |
| Word | 0.5777 | 0.4911 | 0.4991 | 0.6391 | 0.1567 |
| Rouge-W | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6319 | 0.6968 | 0.6787 | 0.6835 | 0.6778 |
| Puertot_1b | 0.6765 | 0.6828 | 0.6973 | 0.6880 | 0.6706 |
| Puertot_2a | 0.6873 | 0.6032 | 0.6921 | 0.6911 | 0.6835 |
| Puertot_2b | 0.6731 | 0.6558 | 0.6671 | 0.6779 | 0.6593 |
| Baseline | 0.1001 | 0.1925 | 0.1697 | 0.1812 | 0.1299 |
| Word | 0.4848 | 0.4016 | 0.4012 | 0.4919 | 0.4178 |

Tabla. 75. Raouge L y Rouge W-Clásico

| Rouge (media) | Rouge-N | Rouge-2 | Rouge-SU4 | Rouge L | Rouge W |
|---|---|---|---|---|---|
| Puertot_1a | 0.6577 | 0.6539 | 0.6625 | 0.6803 | 0.6805 |
| Puertot_1b | 0.6146 | 0.6406 | 0.6695 | 0.6706 | 0.6742 |
| Puertot_2a | 0.6474 | 0.6495 | 0.6405 | 0.6715 | 0.6777 |
| Puertot_2b | 0.6442 | 0.6431 | 0.6470 | 0.6710 | 0.6720 |
| Baseline | 0.2147 | 0.2140 | 0.1973 | 0.1513 | 0.1604 |
| Word | 0.2241 | 0.4720 | 0.4711 | 0.4746 | 0.4561 |

Tabla. 76. Rouge Media

Los resultados del cálculo de la métrica ROUGE de la manera tradicional, obligan a la formulación de una segunda posición de cálculo, mediante una variante denominada Jackknife, pues según D´Cunha (D'Cunha, 2006) es una variante propuesta por la Document Conference Unión (DUC), dicha variante permite obtener una valoración más confiable entre los resúmenes generados por máquinas y los realizados por los humanos. La aplicación de ROUGE con Jackknife obliga a que cada resumen hecho por los humanos (resumen de referencia) sea suprimido de la lista de resúmenes de referencia y añadido a la lista de resúmenes candidatos. Gracias a esta variación en la organización del experimento se logra que los resúmenes de los expertos sean comparados con los extractos generados a partir de sistemas automáticos empleando la misma métrica. Para esto aplicamos de nuevo R-N, ROUGE-2, ROUGE-SU-4, Rouge –L, Rouge W (Ver figura 77 - 82).

| Rouge-N | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6191 | 0.6234 | 0.6891 | 0.6932 | 0.6941 |
| Puertot_1b | 0.6381 | 0.6234 | 0.6591 | 0.6717 | 0.6591 |
| Puertot_2a | 0.6493 | 0.6129 | 0.6654 | 0.6692 | 0.6513 |
| Puertot_2b | 0.6222 | 0.6321 | 0.6647 | 0.6451 | 0.6443 |
| Baseline | 0.1055 | 0.1245 | 0.1801 | 0.1919 | 0.6255 |
| Word | 0.4127 | 0.4321 | 0.4971 | 0.5093 | 0.6917 |
| Rouge-N | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6859 | 0.6632 | 0.6218 | 0.6395 | 0.6598 |
| Puertot_1b | 0.6163 | 0.6998 | 0.6443 | 0.6116 | 0.6156 |
| Puertot_2a | 0.6512 | 0.6591 | 0.6811 | 0.6791 | 0.6975 |
| Puertot_2b | 0.6921 | 0.6878 | 0.6931 | 0.6539 | 0.6923 |

| | | | | | |
|---|---|---|---|---|---|
| Baseline | 0.1583 | 0.1521 | 0.2091 | 0.1292 | 0.1399 |
| Word | 0.4976 | 0.4365 | 0.5722 | 0.4565 | 0.4998 |

Tabla. 77. Rouge N Jackknife

| Rouge-2 | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6291 | 0.6724 | 0.6617 | 0.6773 | 0.6921 |
| Puertot_1b | 0.6518 | 0.6224 | 0.6448 | 0.6891 | 0.6376 |
| Puertot_2a | 0.6773 | 0.6129 | 0.6278 | 0.6792 | 0.6533 |
| Puertot_2b | 0.6111 | 0.6341 | 0.6567 | 0.6819 | 0.6693 |
| Baseline | 0.1535 | 0.1233 | 0.1891 | 0.1729 | 0.6755 |
| Word | 0.4657 | 0.4344 | 0.4851 | 0.4890 | 0.6297 |
| Rouge-2 | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6799 | 0.6418 | 0.6290 | 0.6596 | 0.6478 |
| Puertot_1b | 0.6925 | 0.6348 | 0.6381 | 0.6166 | 0.6296 |
| Puertot_2a | 0.6713 | 0.6441 | 0.6551 | 0.6791 | 0.6765 |
| Puertot_2b | 0.6771 | 0.6458 | 0.6721 | 0.6284 | 0.6533 |
| Baseline | 0.1970 | 0.1223 | 0.1890 | 0.1982 | 0.1379 |
| Word | 0.4457 | 0.4115 | 0.4702 | 0.4519 | 0.4568 |

Tabla. 78. Jackknife

| Rouge/SU4 | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6701 | 0.6724 | 0.6731 | 0.6752 | 0.6731 |
| Puertot_1b | 0.6714 | 0.6224 | 0.6431 | 0.6801 | 0.6356 |
| Puertot_2a | 0.6893 | 0.6159 | 0.6244 | 0.6662 | 0.6573 |
| Puertot_2b | 0.6341 | 0.6311 | 0.6227 | 0.6221 | 0.6633 |
| Baseline | 0.1355 | 0.1235 | 0.1881 | 0.1779 | 0.1735 |
| Word | 0.4577 | 0.4311 | 0.4841 | 0.4880 | 0.4257 |
| Rouge /SU4 | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6599 | 0.6268 | 0.6580 | 0.6635 | 0.6538 |
| Puertot_1b | 0.6665 | 0.6528 | 0.6653 | 0.8126 | 0.6456 |
| Puertot_2a | 0.6133 | 0.6132 | 0.6511 | 0.6121 | 0.6625 |
| Puertot_2b | 0.6671 | 0.6648 | 0.6411 | 0.6829 | 0.6413 |
| Baseline | 0.1971 | 0.1423 | 0.1090 | 0.1182 | 0.1079 |
| Word | 0.4448 | 0.4515 | 0.4902 | 0.4219 | 0.4168 |

Tabla. 79. Rouge SU4 Jackknife

| Rouge-L | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6533 | 0.6914 | 0.6891 | 0.6672 | 0.6931 |
| Puertot_1b | 0.6245 | 0.6234 | 0.6441 | 0.6991 | 0.6796 |
| Puertot_2a | 0.6734 | 0.6979 | 0.6844 | 0.6872 | 0.6893 |
| Puertot_2b | 0.6413 | 0.6812 | 0.6517 | 0.6611 | 0.6183 |
| Baseline | 0.1856 | 0.1625 | 0.1781 | 0.1979 | 0.1865 |
| Word | 0.4817 | 0.4623 | 0.4891 | 0.1581 | 0.1527 |
| Rouge-W | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6239 | 0.6898 | 0.6897 | 0.6975 | 0.6708 |
| Puertot_1b | 0.6825 | 0.6888 | 0.6593 | 0.6697 | 0.6726 |
| Puertot_2a | 0.6823 | 0.6142 | 0.6623 | 0.6515 | 0.6865 |
| Puertot_2b | 0.6811 | 0.6498 | 0.6791 | 0.6789 | 0.6513 |
| Baseline | 0.1091 | 0.1995 | 0.1787 | 0.1823 | 0.1299 |
| Word | 0.5858 | 0.4186 | 0.4552 | 0.4999 | 0.4298 |

Tabla. 80. Rouge L

| Rouge-W | Tex 1 | Tex 2 | Tex 3 | Tex 4 | Tex 5 |
|---|---|---|---|---|---|
| Puertot_1a | 0.6933 | 0.6714 | 0.6791 | 0.6772 | 0.6831 |
| Puertot_1b | 0.6145 | 0.6134 | 0.6541 | 0.6891 | 0.6896 |
| Puertot_2a | 0.6634 | 0.6879 | 0.6744 | 0.6872 | 0.6893 |
| Puertot_2b | 0.6412 | 0.6811 | 0.6527 | 0.6621 | 0.6173 |
| Baseline | 0.1956 | 0.1525 | 0.1681 | 0.1879 | 0.1565 |
| Word | 0.5817 | 0.4621 | 0.4991 | 0.1591 | 0.1517 |
| Rouge-W | Tex 6 | Tex 7 | Tex 8 | Tex 9 | Tex 10 |
| Puertot_1a | 0.6229 | 0.6868 | 0.6887 | 0.6935 | 0.6798 |
| Puertot_1b | 0.6815 | 0.6898 | 0.6573 | 0.6690 | 0.6716 |
| Puertot_2a | 0.6923 | 0.6132 | 0.6621 | 0.6511 | 0.6875 |
| Puertot_2b | 0.6851 | 0.6418 | 0.6771 | 0.6719 | 0.6503 |
| Baseline | 0.1001 | 0.1975 | 0.1797 | 0.1822 | 0.1219 |
| Word | 0.5848 | 0.4136 | 0.4512 | 0.4939 | 0.4198 |

Tabla. . 81. Rouge W

| Rouge (media) | Rouge-N | Rouge-2 | Rouge-SU4 | Rouge L | Rouge W |
|---|---|---|---|---|---|
| Puertot_1a | 0.6589 | 0.6590 | 0.6625 | 0.6775 | 0.6775 |
| Puertot_1b | 0.6439 | 0.6457 | 0.6695 | 0.6643 | 0.6629 |
| Puertot_2a | 0.6616 | 0.6495 | 0.6695 | 0.6729 | 0.6708 |
| Puertot_2b | 0.6627 | 0.6576 | 0.6470 | 0.6593 | 0.6708 |
| Baseline | 0.2016 | 0.2158 | 0.1275 | 0.1710 | 0.1642 |
| Word | 0.2462 | 0.4254 | 0.4511 | 0.4133 | 0.4217 |

Tabla 82. Rouge Media

#### 7.8.3.2.1.- Análisis de los Resultados de la Aplicación de Rouge

Los resultados de la aplicación de Rouge demuestran la calidad de los resúmenes obtenidos mediante las reglas PUERTOTERM. En ambas mediciones, tanto con ROUGE clásico como con Jackknife. Los resúmenes generados con el modelo que se propone en esta investigación superan a los resúmenes baseline y a los obtenidos mediante Microsoft Word (Anexo 33) tanto con ROUGE- N, ROUGE-2, ROUGE-SU-4, ROUGE-L, ROUGE W para ambas mediciones nótese estos valores en las tablas que se muestran a continuación (Ver tablas 72- 82).

Puede observarse que los extractos Puertot_1ª, obtienen la máxima puntuación en casi todas las mediciones, menos ROUGE-SU-4, debido a que las reglas Puertot 1ª se verificaron con oraciones de menor longitud que los otros textos. En calidad le siguen los resúmenes de Puertot 2 a y 2 b, en las que se obtienen valores muy similares con todas las medidas y resultados muy relevantes. Para el final aparecen las reglas puertot 1b, en las que se obtienen buenos resultados, sin embargo la estructura de los textos que sirvieron de base a su análisis no tenían la homogeneidad más alta en cuanto a estructura. En comparación con word y baseline nuestros 4 tipos de resumen obtienen puntuaciones superiores con las 5 medidas que se utilizan para la evaluación. Los resúmenes Word son superiores a los resúmenes baseline que poseen un rango sobre 0.1 (Ver tablas 72-82).

Si observamos los resultados de la aplicación de las métricas con jackknife, podemos ver que los resultados tienden a la similaridad y en algunos casos la similitud en ngramas aumenta, algo que es muy bueno ya que indica que los resúmenes se parecen más a los de los humanos. Los resúmenes puertot_1a obtienen la puntuación máxima (0.6589 con ROUGE-N, 0.6590, ROUGE-2, 0.6625 con Rouge-SU-4, 0.6719 con Rouge-L y 0.6775 con Rouge L). Se corrobora la calidad superior de los 4 tipos de resúmenes del modelo con respecto a baseline y word, siento baseline el más bajo con las calificaciones siguientes: 0.2016 con Rouge-N, 0.2158 con Rouge –2, 0.1275 Rouge-SU4, 0.1545 Rouge-L y 0.1642 para Rouge W.

De los cuatro tipos de resúmenes obtenidos mediante la aplicación del modelo objeto de la investigación, son los puertot_1ª (Anexo 28), los de mayor calidad, pues cargan en sí todas las reglas puertoterm y las reglas de desambiguación léxica. Los resúmenes Puertot_2ª (Anexo 30) y 2b (Anexo 31) son los segundos en calidad los de mayor calidad, pues poseen un mayor número de oraciones 27 y 29 como promedio. Los resúmenes Puertot_1b (Anexo 29), son aquellos en los cuales solo se ha aplicado un modelo de abstracción, son de menor puntuación con respecto a los otros, debido a que se eliminaron elementos importantes del texto. A igual que en el experimento de (D´cuhna, 2005) la baja puntuación que se obtiene parece estar sustentada en la prevalencia en el corpus de análisis de resúmenes de abstracción generados por los expertos.

Los resúmenes baseline y los resúmenes del Word obtienen menos puntuación cuando se les mide con Rouge obtienen en la evaluación con ROUGE, con lo cual pueden ser considerados como los resúmenes de menor calidad.

### *7.9.- Evaluación de la Cohesión y la Coherencia Textual*

Las medidas que se declararon solo son proporciones geométricas, es por ello que autores como Pinto (Pinto, 2001) y Hernández (Hernández, 2007), enfatizan en que los resúmenes se evalúan de acuerdo a métricas de similitud establecida sobre patrones físicos, en este caso los bigramas, por ello en esta

investigación se insiste en la evaluación de la cohesión y la coherencia del texto como única vía de saber si los textos resumidos cumplen con su objetivo pragmático Pinto (Pinto, 2001), resúmenes legibles y entendibles por los humanos. Para realizar estos análisis sometimos a todos los textos que se analizaron en los experimentos a la valoración de 2 jueces, los que determinaron el nivel de cohesión, coherencia y balance de textual de los textos resultantes de la aplicación de los diferentes modelos de resumen que se utilizan en la evaluación (Anexo 15) (Ver figura 140).

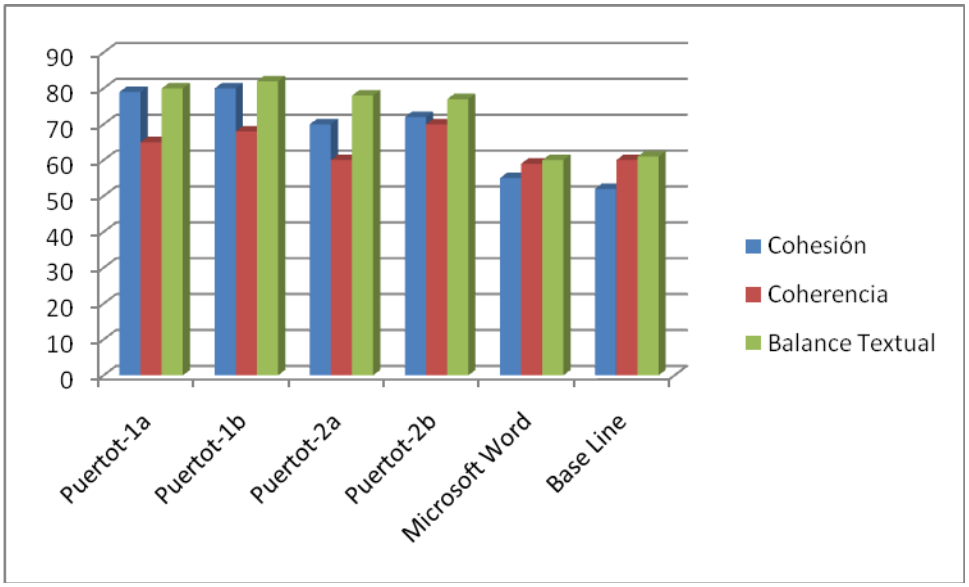



Figura. 140. Evaluación de la Cohesión y la Coherencia Textual

Los resultados de la evaluación de la coherencia, la cohesión y el balance textual evidencian que el modelo propuesto facilita textos de elevado nivel de comprensión, pues presenta textos legibles y comprensibles por las personas que los analizan, lo que evidencia que los métodos basados en el uso de marcas cohesivas en el texto superan a los modelos netamente estadísticos. La media de cohesión se refleja en un 0.7529, debido a que los textos que poseen mayor unidad textual son tienden a ser menos unidos. Desde el punto de vista de la coherencia se obtiene una menor medida, los textos aunque están marcados tienden a disminuir en cohesión debido a que el sentido pragmático de los mismos no puede aun ser Interpretado por un ordenador. El balance textual, que se adecua a la forma de presentar las oraciones y la carga

de los párrafos se evalúa en un 0.7772, algo que evidencia la calidad alcanzada en los objetivos pragmáticos de los textos.

### *7.10.- Evaluación de la usabilidad del Sistema*

La evaluación de un producto o sistema no concluye hasta tanto los usuarios finales de este no hayan emitido los criterios sobre este. En trabajos de diversos expertos se insiste en métodos para evaluar la calidad de los sistemas de Recuperación de la Información entre ellos se encuentran: Johson (Johnson et al., 2003), así como a Marcos y Cañada (Marcos and Gómez, 2006) y Granollers (Granollers et al., 2005), estos últimos enfatizan en el análisis de la interacción del sistema con los usuarios.

Para acometer esta parte del análisis se han tomado como referencia algunas de las obras más destacadas sobre el diseño centrado en el usuario y en particular los principios heurísticos de Jakob Nielsen (Nielsen, 1994, Nielsen, 2002a, Nielsen, 2002b) y Machón (Manchón, 2002). Teniendo en cuenta el tipo de sistema que estamos evaluando, creamos una plantilla que recoge indicadores de análisis relativos a los aspectos siguientes: la navegación, la funcionalidad, el control por parte del usuario, el uso de la lengua, la ayuda en línea y guía del usuario, la información proporcionada por el sistema, la accesibilidad, la coherencia, la prevención y corrección de errores, y la claridad arquitectónica y visual del sistema.

Han sido muchas las formas de evaluar la usabilidad de un sistema. En esta investigación se ha determinado usar un estudio de usabilidad denominado usabilidad de"inspección" una forma de evaluación que tiene nexos directos con la evaluación heurística, actividad que consiste en el concurso de un grupo de evaluadores a los que se encarga la valoración de la interfaz del sistema mediante principios heurísticos de usabilidad. Cada evaluador realiza la valoración de forma individual como si fuese el usuario del sistema. La revisión se realiza de manera individual y asumiendo el papel de usuario. Cuando termina el último evaluador se divulgan los resultados ya analizados.

Para lograr una uniformidad en le evaluación del sistema el autor desarrolló un cuestionario (Anexo 9) donde se evalúan las variables que se utilizan en la investigación heurística. En el referido cuestionario se le dio a los evaluadores tres opciones: "sí, siempre", "no, nunca" y "a veces", a las que se le dio el valor de O, 1 y 2, siendo 2 el valor más alto. Cuando los expertos ofrecieron sus respuestas, se tabularon las afirmativas, las negativas y las indeterminadas ("a veces"). Con este proceder las respuestas de los evaluadores obtienen cualidades numéricas que se utilizan para valorar los elementos del sistema.

Para la evaluación del sistema Puertotex se usaron 5 expertos (Especialistas) en la materia y 42 usuarios (Comunidad de Ciencias Biológicas) dentro del experimento, ya que el sistema no solo brinda información a expertos, sino también a personal que usa la información para otras tareas, que no son únicamente de investigación.

En la segunda fase de la evaluación, después tener la valoración de los jueces (los 5 expertos) se aplica este test a usuarios foráneos del sistema con el fin de determinar otros errores que no se hayan obtenido en la prueba con los evaluadores sobre al sistema PUERTOTEX para realizar con valoraciones (Anexo 10). Los parámetros que se evalúan tienen como centro la heurística propuesta por Jacob Nielsen:

1. **Navegación:** El Culmen de información en la memoria del usuario en el momento de usar el sistema, aspecto que necesita ser mínimo. Desde el punto de vista heurístico son más sólidos los procesos de deducción o reconocimiento que los de memoria. Un sistema de navegación de calidad facilita la recuperación de la información de forma óptima.
2. **Funcionalidad:** Todos los parámetros de utilización del sistema deben recogerse en el segmento de descripción del mismo es por ello que debe aparecer descrito en el sistema: lenguas en que trabaja, prestaciones adicionales, cuales son los usuarios que interactúan con él, tema que contiene el sistema, etc.

3. **Control del usuario:** En muchos sistemas el usuario parece temerle a la herramienta de trabajo y que el sistema mismo es el que controla sus acciones. El usuario es quien debe guiar el sistema, sentir que la herramienta responde a sus órdenes que tiene libertad para moverse por ella. Es vital en este aspecto hacer flexible la interfaz de usuario para que este recupere mejor la información que desea.
4. **Lenguaje y contenido:** Los textos que se ponen en el sistema deben ser claros y comprensibles para los usuarios, ya que los usuarios que usan estas aplicaciones no son esencialmente expertos en el tema.
5. **Ayuda en línea:** El sistema debe incorporar mecanismos para reconocer, diagnosticar y solucionar errores.
6. **Información del sistema:** Al usuario hay que diferenciarle las interfaces, tanto la de consulta o la de resultados y también debe saber la posibilidad para ejecutar acciones.
7. **Accesibilidad:** El sistema debe plantearse cuestiones básicas para determinados usuarios entre ellos para usuarios que poseen discapacidad física o poca alfabetización tecnológica.
8. **Coherencia:** La estructura y el diseño de las páginas del sistema debe ser siempre uniforme para todas sus páginas.
9. **Prevención errores:** Un sistema que no está preparado para evitar los errores de sus futuros usuarios, es un sistema frágil.
10. **Claridad arquitectónica:** La interfaz del usuario debe ser clara de modo que facilite la localización de la información.

En esta tabla se muestran los resultados de la aplicación del test de heurístico del uso del usuario. La navegación y la claridad arquitectónica son los elementos que menor nivel alcanzan debido a que los usuarios no están adaptados a trabajar con sistemas que utilicen Web semántica y ontologías como herramientas y en la mayoría de los casos no logran identificar a qué lugares del sistema dirigir su búsqueda. El sistema es funcional, pues un 84 % de los encuestados considera ve la presencia en él de diversas informaciones relativas a idioma, lenguas, esquema de usuarios, algoritmos de trabajo, etc. Como todo sistema de información PUERTOTEX logra que sus usuarios se

sientan seguros y con control de las acciones, hay usuarios que no poseen un entrenamiento con estos sistemas y en algunas ocasiones sienten sentirse espiados por el sistema, diciendo en ocasiones expresiones como “este sistema piensa por mí”, y es en este momento donde entran en pánico. Un 100 % de los evaluadores dejó bien claro que el sistema muestra información relativa a idioma y contenidos. El 85 % de los usuarios del sistema confiesa que el sistema en accesible siempre y solo el 18 % a veces. La coherencia del diseño de las páginas Web fue vista por el 90 % de los encuestados. Solo el 60 % de los casos ve que el sistema tenga un sistema de prevención de errores y el 65 % aprecia en él claridad arquitectónica, aspectos en los que en otras versiones habrá que trabajar (Ver Tabla 83).

| Variables de Uso | Sí | A veces | *No* |
|---|---|---|---|
| Navegación | 62,0 | 18,0 | 10 |
| Funcionalidad | 81,4 | 10,6 | 18 |
| Control del usuario | 89,0 | 11,0 | 0 |
| Lengua y contenido | 100 | 0 | 0 |
| Ayuda en línea | 100 | 0 | 0 |
| Información del sistema | 100 | 0 | 0 |
| Accesibilidad | 85,0 | 15,0 | 0 |
| Coherencia | 90,0 | 10,0 | 0 |
| Prevención errores | 60,0 | 20,0 | 20 |
| Claridad arquitectónica | 65,5 | 25 | 10,8 |

Tabla 83. Usabilidad del Sistema

Para concluir la investigación se aplicó un test de usabilidad a los usuarios y los expertos, a partir de un conjunto de tareas específicas con Puertotex. Se preparó una serie de 8 tareas y se realizó la prueba con 10 usuarios y 5 expertos, de manera individual, durante 40 minutos. Las tareas que se acometieron fueron las siguientes y luego la media de cada una de ellas se contabilizó y se puso en una tabla (Ver Tabla 84).

| Tarea | Tipo de tarea | Puntuación |
|---|---|---|
| 1 | Búsqueda de un término en dos idiomas | 99,55 |
| 3 | Búsqueda de información no lingüística | 96,85 |
| 1 | Búsqueda de lema | 100 |
| 1 | Claridad de presentación de la información (idiomas) | 100 |
| 2 | Búsqueda de un término | 98,00 |
| 4 | Feedback | 95,00 |
| 5 | Claridad de resultados | 86,00 |
| 1 | Búsqueda de Resúmenes | 100,00 |

Tabla 84. Tarea de Estudio

La mayoría de los parámetros que se analizan en la evaluación de usabilidad son favorables, la Búsqueda de un término en dos idiomas arrojó un 99, 5 de efectividad. La recuperación de información no lingüística un 96.5 %. La recuperación y búsqueda de información lematizada fue efectiva en un 100 %, así como la claridad de la Representación de la Información. La búsqueda de un término es efectiva en un 100 % al igual que la búsqueda de resúmenes. Un aspecto al que debe prestar atención el sistema es la presentación con claridad de los resultados.

### *7.11.-Valoración del Sistema PUERTOTEX*

El sistema que se ha propuesto ha servido para la validación de los resultados del Metamodelo para la Extracción y Desambiguación de Textos. Los procederes desarrollados para la construcción del sistema muestran la validez y el rigor que se tuvo en cada etapa de desarrollo del software. Las pruebas y los test aquí expuestos son solo una parte de los procesos que deben desarrollarse cuando se piensa hacer un sistema basado en Minería Textual. La evaluación del sistema, aunque extremadamente compleja por la intercalación de técnicas cualitativas, cuantitativas y el desarrollo de modelos de evaluación propios para el sistema ha mostrado la efectividad de los postulados teóricos de la investigación. El estudio analítico de los corpus, por

técnicas de lexicometría y análisis de discurso revela las condiciones de la estructura retórica del texto y los vocablos recurrentes dentro del dominio que se estudia, esto permitió seleccionar el texto del sistema y establecer cuál de los dos era más difícil de automatizar, en este caso el español. Los estudios de agrupamiento y valoración completaron los resultados de las técnicas lexicométricas y estudio de discurso del capítulo 5 y de conjunto con técnicas de clustering que se ofrecen en este apartado 7 validaron con modelos heurísticos la calidad de los textos para el sistema. La ontología del sistema constituía un gran problema, un sistema con un modelo de estructuración desarrollado para prestaciones complejas dentro de la Minería de Texto, para ello fue necesario desarrollar una Metodología particular que permitiese la valoración de la ontología Ontosatcol en todas sus variedades, esto permitió reajustar la ontología antes que el sistema se pusiera en marcha de los contrario los errores operacionales del mismo non podrían ser salvados. La evaluación con 5 variantes de Rouge, da una confiabilidad en la calidad de los resúmenes obtenidos, esta evaluación se realizó después con Rouge N grama para determinar la calidad de los vocablos en la recuperación de la información con resultados satisfactorios. Finalmente Estos resultados permitieron una valoración exacta de las cualidades lingüísticas de los corpus.

### *7.12.- Valoración de Especialistas*

Como en toda investigación que propone un modelo el final de la evaluación concluye con la evaluación del modelo por criterio de expertos. En nuestro caso se utilizó como referente la forma de evaluación que propone Crespo (Crespo, 2007) la misma tiene como premisa la asignación de valor a los expertos tomando como base sus fuentes de argumentación. La selección del personal experto fue llevada a cabo de la siguiente manera:

- Focalización de los posibles expertos partiendo de la base del conocimiento que posee el doctorando.

- Identificación de los expertos a partir de las competencias, declaradas en las respuestas al instrumento de investigación (cuestionario), donde exponen sus juicios sobre el fenómeno objeto de estudio.
- Obtención del coeficiente K, teniendo en cuenta la postura que asumen los encuestados en lo referente a su nivel de conocimiento del tema que se investiga y mediante de las fuentes que sustenten teóricamente sus respuestas (Anexo 34).
- Selección de Expertos: Segunda Fase de la Técnica: Se contó con 11 expertos en total. La tabla (Tabla 85) muestra su distribución general.
- Los expertos seleccionados contestaron otro cuestionario en el cual valoraron los principios del modelo en una escala de *Muy importante, Bastante importante, Importante, Poco importante, No importante* (Anexo 35).
- La Funcionalidad del modelo fue valorada en una escala 5 estados donde 5 (Imprescindible para lograr la funcionalidad del modelo); 4(Muy útil para lograr la funcionalidad del modelo); 3 (Útil para lograr la funcionalidad del modelo); 2 (Quizás podría servir para lograr la funcionalidad del modelo); 1- No aporta nada a la funcionalidad del modelo (Tabla 85) (Anexo 35).

| Expertos | Centro de Procedencia | Años de Experiencia en Procesamiento de la Información |
|---|---|---|
| 1 | Universidad de la Habana | 30 |
| 2 | Universidad de la Habana | 31 |
| 3 | Universidad de Granada | 20 |
| 4 | Universidad de Granada | 25 |
| 5 | Universidad de Granada | 30 |
| 6 | Universidad Simón Bolívar Venezuela | 12 |
| 7 | Universidad Central de las Villas, Cuba | 40 |
| 8 | Universidad Central de las Villas, Cuba | 12 |
| 9 | Universidad Central de las Villas, Cuba | 18 |
| 10 | Instituto Internacional de Periodismo | 16 |
| 11 | Biblioteca de la AECID, España | 8 |

Tabla 85. Procedencia de los Expertos que Evalúan el Modelo

La mayoría de las opiniones de los expertos se declaran a continuación:

- El 95 % de los expertos considera como *muy importante* generar modelos para la construcción de resúmenes automáticos con los presupuestos de la Ciencia de la Información.
- 2 expertos consideran poco necesarios los estudios de necesidades de información.
- El 100 % de los que evalúan el modelo coincide en que la holística y la sinergia del modelo es vital para alcanzar resultados de calidad.
- Todos los evaluadores opinan que es muy importante mezclar diversas disciplinas en el desarrollo del modelo.
- Los expertos consideran como muy importante el estudio de estrategias cognitivas y el análisis de dominio.
- El 90 % piensa que es muy importante construir reglas heurísticas a partir de las regularidades del discurso del dominio. Un 10 % opina que eso es poco importante.
- El 85 % de los encuestados refiere que es bastante importante realizar la evaluación de los resúmenes por técnicas estadísticas y la evaluación de jueces, sin embargo un 15 % lo considera este aspecto muy importante.
- Evaluar el sistema de resúmenes ha sido un aspecto que ha sido considerado como muy importante por los expertos
- Un experto considera que *no es importante* formular una evaluación integral del modelo de resumen, sin embargo el 90% lo cataloga en las categorías de *muy importante* y *bastante importante*.
- 2 evaluadores consideran que no es necesario trabajar con ontologías en el sistema, sin embargo los otros 9 opinan que este proceso está entre *muy importante* y *bastante importante*
- Construir resúmenes automáticos sobre documentos científicos ha sido valorado como un proceso muy importante por 10 usuarios, solo uno lo considera bastante importante.

- 85 % de los expertos responde entre muy importante y bastante importante cuando se le pregunta sobre los roles de los diversos actores en el sistema.

- El 95 % de los especialistas cree que es muy importante que el modelo vaya en búsqueda de la calidad de los resúmenes, solo un 5 % cree que es bastante importante esta labor.
- La visualización de la información ha sido valorada como muy importante por 10 expertos, solo uno opina que es bastante importante.
- El 100 % expresa que es muy importante la capacidad que tiene el modelo de describir recursos documentales utilizando estándares.
- La Búsqueda de información a partir de la semántica que declaren los conceptos obtiene la categoría de muy importante por el 100 % de los expertos.

- El 95 % de los expertos opina que es bastante importante el desarrollo de redes sociales, el 5 % cree que esto este aspecto no es poco importante.


# Referencias Bibliográficas

## *Referencias Bibliográficas*

ANSI 1978. *American National Standard for writing abastracts,* NuevaYork, ANSI.

ARCO, L. 2005. *Corpus Miner.* Tesis de Maestría, Universidad Central "Marta Abreu" de las Villas.

ARCO, L. 2007. *Corpus miner: herramienta para el etiquetado de grupos y la obtención de extractos.* MSc., Universidad Central de las Villas.

ARCO, L. 2008. *Agrupamiento basado en intermediación diferencial.* Tesis Doctoral, Universidad Central"Marta Abreu" de las Villas.

BOEHM, B. W. 1981. *Software Engineering Economics.*

BOUILLON, P., CLAVEAU, V., FABRE, C. & SEBILLOTE, P. n.d. Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method.

BRANK, J., GROBELNIK, M. & MLADENIĆ, D. 2005. A survey of ontology evaluation techniques. *SIKDD 2005 at multiconference IS 2005.* Ljubljana, Slovenia.

BREWSTER, C., ALANI, H., DASMAHAPATRA, S. & WILKS, Y. 2004. Data Driven Ontology Evaluation. *Proceedings of the Language Resources and Evaluation Conference (LREC 2004).* Lisboa, Portugal.

CARROLL, J., MINNEN, G. & BRISCOE, T. 1999. Corpus Annotation for Parser Evaluation. *EACL workshop on Linguistically Interpreted Corpora (LINC).* Bergen, Norway.

CRESPO, T. 2007. *Respuestas a 16 preguntas sobre el empleo de expertos en la investigación pedagógica,* Perú, San Marcos.

CUNHA, I. D., FERNÁNDEZ, S. & VELÁZQUEZ MORALES, P. 2007. A New Hybrid Summarizer Based on Vector Space Model, Statistical Physics and Linguistics. *In:* GELBUKH, A. & KURI MORALES, F. (eds.) *MICAI 2007.* Berlín: Springer-Verlag Berlin Heidelberg.

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DOMÍGUEZ, S. 2010. PROTEX. beta ed. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de automática.

DOMÍGUEZ, S. 2011. FOXCORP. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de Ingeniería Automática.

DOMÍNGUEZ, S. 2010. PROTEX. beta ed. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de automática.

DOMÍNGUEZ, S. 2011. FOXCORP. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de Ingeniería Automática.

EVERMANN, J. E. & FANG, J. 2010. Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems,* 35**,** 391-403.

FAHAD, M. & ABDUL QADIR, M. 2008. A Framework for Ontology Evaluation. *Supplementary Proceedings of the 16th International Conference on Conceptual Structures (ICCS'08).* Toulouse, France.

FELLBAUM, C. 1998. *WordNet - An Electronic Lexical Database*.

FERNÁNDEZ-BREIS, J. T., EGAÑA ARANGUREN, M. & STEVENS, R. 2009. Quality evaluation framework for bio-ontologies. *In ICBO: International Conference on Biomedical Ontology, 2009* Buffalo, New York.

FERNÁNDEZ-LÓPEZ, M., GÓMEZ-PÉREZ, A. & JURISTO, N. 1997. Methontology: from ontological art towarads ontological engineering. *In Spring Symposium on Ontological. Engineering of AAAI, 1997.* California. : Stanford University.

FERNÁNDEZ, O., TORAL, A. & MUÑOZ, R. 2009. Exploiting Lexical Measures and a Semantic LR to Tackle Textual Entailment in Italian.

FRAKES, W. & BAEZA-YATES, R. 1992. *Information Retrieval :data Structure & Algorithms,* New York

GANGEMI, A., CATENACCI, C., CIARAMITA, M. & LEHMANN, J. 2005. A Theoretical Framework for Ontology Evaluation and Validation.

GÓMEZ PÉREZ, A. 1994. *Some Ideas and Examples to Evaluate Ontologies*, Knowledge Systems Laboratory, Stanford University

GRANOLLERS, T., LORÉS, J. & CAÑAS, J. J. 2005. *Diseño de sistemas interactivos centrados en el usuario,* Barcelona, UOC.

GRUNNINGER, M. & FOX, M. Year. Methodology for the Design and Evaluation of Ontologies. *In:* In: Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, 1995 Montreal.

GUARINO, N. & WELTY, C. 2002. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM,* 45**,** 61-65.

GUPTA, S. & STENT, A. J. n.d. Automatic Evaluation of Referring Expression Generation Using Corpora.

HARTMANN, J., SPYNS, P., GIBOIN, A., MAYNARD, D., CUEL, R., SUÁREZ-FIGUEROA, M. C. & SURE, Y. 2005. *D1.2.3 Methods for ontology evaluation*, Knowledgeweb.

HERDAN, G. 1960. *Type-Token Mathematics: A Textbook of Mathematical Linguistics,* The Hague, The Netherlands, The Netherlands: Mouton & Co.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

JAAN KAALEP, H. & VESKIS, K. 2007. Comparing Parallel Corpora and Evaluating their Quality.

JOHNSON, F. C., GRIFFITHS, J. R. & HARTLEY, R. J. 2003. Task dimensions of user evaluations of information retrieval systems. *nformation Retrieval,* 18.

KENT, A. 1955. Machine literature searching VIII. Operational Criteria for Designing Information Retrieval Systems. *American Documentation,* 6**,** 93-101.

KILGARIFF, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics,,* 6.

KOEHN, P. n.d. Europarl: a multilingua corpus for evaluation of machine translation.

KUHN, T. 1979a. *The structure of scientific revolutions,* Chicago, University of Chicago Press.

KUHN, T. 1979b. *The structure of scientific revolutions,,* Chicago, University of Chicago Press.

LANQUILLON, C. 2002. Enhancing Text Classification to Improve Information Filtering. *Künstliche Intelligenz***,** 37-38.

LEIVA, A., SENSO, J., DOMÍNGUEZ, S. & HÍPOLA, P. 2009. An Automat for the semantic processing of structured information. *In ISDA 9na International Conference of Desing of Software and Aplicactión.* Italia, Pissa: IEEE.

LEWIS, D. D. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *15th Ann Int'1 SIGIR '92.* Denmark.

LIN, C. & HOVY, E. 1998. Automatic Evaluation of summaries using n-gram co-ocurrence Statistic. *In Proceeding of HLTNAACL.* EE.UU.

LUNH, H. 1958. The Automatic creation of Literature abstracts. *Journal of Research of Development***,** 159 – 165.

LOZANO-TELLO, A. & GÓMEZ-PÉREZ, A. 2004. ONTOMETRIC: A method to choose the appropriate ontology. . *Journal of Database Management,* 15.

MAEDCHE, A., MOTIK, B., STOJANOVIC, L., STUDER, R. & VOLZ, R. 2003. Ontologies for Enterprise Knowledge Management. *EEE Intelligent System***,** 26-34.

MAEDCHE, A. & STAAB, S. Year. Measuring Similarity between Ontologies. *In:* In: Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002, 2002a Madrid, Spain. NCS/LNAI 2473, Springer.

MAEDCHE, A. & STAAB, S. 2002b. Measuring Similarity between Ontologies. *In: Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002, NCS/LNAI 2473.* Madrid, Spain.: Springer.

MANCHÓN, E. 2002. *Evaluación por criterios o heurística* [Online]. Available: http://www.ainda.info/evaluacion_heuristica.html [Accessed 1.enero 2011].

MARCOS, M. C. & GÓMEZ, M. 2006. Idoneidad de las interfaces de léxicos y terminologías en la web : Glat: Aspects méthodologiques pour l'élaboration de lexiques unilingues et de multilingues. *Bertinoro*, 17-20.

MCQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. pp.182-297.

MIHALCEA, R., CORLEY, C. & STRAPPARAVA, C. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity

NIELSEN, J. 1994. Heuristic evaluation. *In:* NIELSEN, J. & MACK, R. (eds.) *Usability Inspection Methods.* New York, NY.: John Wiley & Sons,.

NIELSEN, J. 2002a. *How to Conduct a Heuristic Evaluation* [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_evaluation.html [Accessed 26. enero 2011].

NIELSEN, J. 2002b. *Ten Usability Heuristics* [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_list.html [Accessed 21.enero 2011].

NING, H. & SHIHAN, D. Year. Structure-Based Ontology Evaluation. *In:* International Conference on e-business on Engineering, 2006a. Computer Society, IEEE.

NING, H. & SHIHAN, D. 2006b. Structure-Based Ontology Evaluation. *In: International Conference on e-business on Engineering.* Computer Society, IEEE.

OBRST, L., WERNER, C., INDERJEET, M., RAY, S. & SMITH, B. 2007. The Evaluation of Ontologies: Toward Improved Semantic Interoperability. *In:* BAKER, C. J. O. & CHEUNG, K.-H. (eds.) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences.* Springer.

ONCINIZ-MARTÍNEZ, J. L. 2009. Towards a Corpus-Based Analysis of Anglicisms in Spanish: A Case Study. *International Journal of English Studies*, 115-132.

PARSONS, K., MCCORMAC, A. & BUTAVICIUS, M. 2009. *Human Dimensions of Corpora Comparison: an Analysis of Kilgarriff's (2001) Approach,* Melbourne, Command, Control, Communications and Intelligence Division Defence Science and Technology Organisation.

PASLARU BONTAS, E. & MOCHOL, M. 2005. A Cost Model for Ontology Engineering. *In:* BERLIN, F. U. (ed.).

PINEDA, L. A., CASTELLANOS, H., CUÉTARAB, J., GALESCU, L., JUÁREZ, J., LLISTERRID, J., PÉREZA, P. & VILLASEÑORE, L. 2002. The Corpus DIMEx100: Transcription and Evaluation.

PINTO AVEDAÑO, D. E. 2008. *On Clustering and Evaluation of Narrow Domain Short-Text Corpora.* PHD, Universidad Polit´ecnica de Valencia, Departamento de Sistemas Informáticos y Computación, Reconocimiento de Formas e Inteligencia Artificial.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez

QADIR, M. & NOSHAIRWAN, W. 2007a. Las ontologías: advertencias para la omisión del Conocimiento en disyuntas Las ontologías. *Segunda Conferencia Internacional sobre Internet y sus aplicaciones Web y Servicios (ICIW07), 2007.*: . IEEE.

QADIR, M. & NOSHAIRWAN, W. 2007b. Warnings for Disjoint Knowledge Omission in Ontologies. *Second International Conference on internet and Web Applications and Services (ICIW07).* IEEE.

RAILEANU, D., BUITELAAR, P., VINTAR, S. & BAY, J. 1999. Evaluation Corpora for Sense Disambiguation in the Medical Domain.

RAJAN, K., RAMALINGAM, V., GANESAN, M., PALANIVEL, S. & PALANIAPPAN, B. 2009. Automatic classification of Tamil documents using vector space model and artificial neural network. *Expert Systems with Applications,* 36, 10914–10918.

RAMOS, E., NÚÑEZ, H. & CASAÑAS, R. 2009. Esquemas para evaluar ontologías únicas para un dominio de conocimiento. *Revista Venezolana de Información, Tecnología y Cococimiento,* 6**,** 57-71.

RIJSBERGEN, C. J. 1979. *Information Retrieval. ,* London, Butterworths.

ROBERTS, A., GAIZAUSKAS, R., HEPPLE, M., DEMETRIOU, G., GUO, Y., ROBERTS, I. & SETZER, A. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics,* 42**,** 950–966.

SABOU, M., LÓPEZ, V., MOTTA, E. & UREN, V. 2006. Ontology Selection:Ontology Evaluation on the Real Semantic Web. *WWW2006.* Edimburgo.

SAGGION, H. & LAPALME, G. 2002. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics* 28**,** 497-526.

SALTON, G. & MCGILLM, M. 1983. *Introduction to modern information retrieval.,* Nueva York:, McGraw-Hill.

SALTON, G., WONG, A. & YANG, C. S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,,* 18**,** 613-620.

SHAMS, R. & ELSAYED, A. 2008. A Corpus-based Evaluation of Lexical Components of a Domainspecific Text to Knowledge Mapping Prototype. *11th International Conference on Computer and Information Technology (ICCIT 2008).* Khulna, Bangladesh.

SMITH, T. F. & WATERMAN, M. S. 1981. Identication of common molecular subsequences. *Journal of Molecular Biology,* 147**,** 195-197.

SPYNS, P. 2005. EvaLexon: Assessing triples mined from texts. Bruselas.

STAAB, S. 2004. Why Evaluate Ontology Technologies? Because It Works! *IEEE INTELLIGENT SYSTEMS*.

STEIN, G., BAGGA, A. & WISE, G. B. 2000. Multi-document summarization: Methodologies and evaluations. *TALN.*

UNE 1990. *Documentación : preparación de resúmenes 50-103-90* Madrid, AENOR.

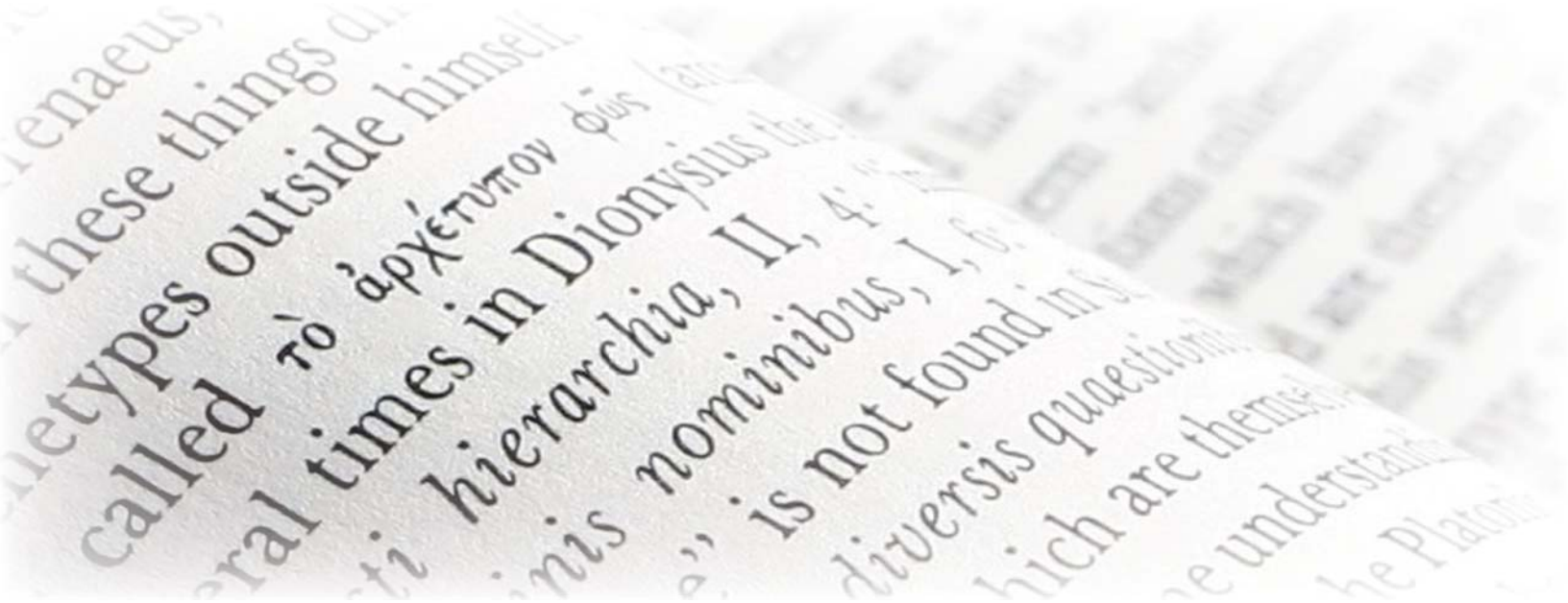
USCHOLD, M. & KING, M. 1995. Towards a Methodology for Building Ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing*.

WAYNE, C. n.d. Multilingual Topic Detection and Tracking:Successful Research Enabled by Corpora and Evaluation

YAO, H., MARK, A., ORME, M. & ETZKORN, L. 2005. Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science,,* 1**,** 107-113.

YU, J., JAMES, A. T. & TAM, A. 2009. Requirements-oriented methodology for evaluating ontologies. *Information Systems Research,* 34**,** 766-791.

ZHANG, T., XU, D. & CHEN, J. 2008. Application-oriented purely semantic precision and recall for ontology mapping evaluation. . *Knowledge-Based Systems, ,* 21**,** 794-799.




# Capítulo VIII
## Conclusiones

# 8.- Conclusiones

- La confección de resúmenes por vía automática es un proceso gestado en el terreno de la Inteligencia Artificial en la década del 60 de pasado siglo, sin embargo, las posiciones teóricas que ha venido desarrollando esta ciencia están en franca fase de estancamiento debido a la inexistencia en sus postulados de elementos diversos como: factores cognitivos, lingüísticos, semánticos y semióticos, los que hacen del proceso de resumen algo más que una suma compleja de algoritmos. Por otra parte, los estudios teóricos que existen en esta área son esgrimidos casi en su totalidad desde la Ciencia de la Computación, lo que limita el alcance teórico de los supuestos de este tema, ya que esta disciplina por sí sola no puede lograr la imbricación de los saberes que se demandan para la construcción de un supuesto teórico que sea capaz de tener en cuenta todos los cambios que han ocurrido desde el siglo XX hasta la actualidad en los receptores y productores de la información.
- Son insuficientes los elementos que deben ser utilizados en un sistema de resúmenes. La mayoría de los estudios identifican como elementos necesarios para la confección del resumen automático la existencia de un corpus para ajustar en él estrategias de PLN (Procesamiento del Lenguaje Natural). Sin embargo, la praxis demuestra que los elementos que tienen que estar presentes en un sistema de resumen son los corpus anotados, los diccionarios semánticos, el estudio de la comunidad epistémica a la que se dedica el producto, el desarrollo de bases de conocimiento, así como las técnicas de marcado de textos y la detección de tópicos, estos elementos indiscutiblemente están en función de lograr un resumen que pueda ser leído. Para llevar a buen puerto estas demandas en el resumen automático se necesita de miradas teóricas y metodológicas hacia el terreno de las competencias informativas y representacionales, donde el análisis de domino y los entornos comunicacionales tengan una preponderancia.
- En el campo resumen automático poco puede decirse en lo referente a normas. La normalización no es un terreno fértil en esta área que aún está distante de lograr resultados de calidad que indiquen la existencia de

procedimientos estándar para la sumarización de textos. Son muy conocidas las normas que se encargan de regir los procesos de resumen por medios tradicionales, sin embargo, en la minería textual más que de normas, habría que hablar de procedimientos generales que permitiesen desarrollar resúmenes bajo determinadas condiciones, algo que tampoco se vislumbra a corto plazo, debido a la poca operatividad de la mayoría de los procesos de construcción y a la elevada proliferación de técnicas de construcción automática que conducen en la mayoría de los casos a un mismo resultado.

- Las técnicas de construcción automática de texto están formuladas a partir de la dimensión humana del proceso, tratando de calcar las estrategias cognitivas de los humanos en el acto del resumen, sin embargo hay factores de índole perceptivo, semiológico, situacional, motivacional y cognitivo, donde las técnicas que ha desarrollado la Ciencia de la Computación no han logrado penetrar, ya que para el desarrollo de herramientas y procedimientos anclados en estos presupuestos necesitan del abordaje de un paradigma multidisciplinar que les permita concebir sus instrumentos metodológicos (netamente matemáticas) como instrumentos de estructura múltiple.
- Los software que hoy realizan la minería de texto y la extracción automática de corpus textuales han sido formulados desde la visión de la Ciencia de la Computación y la Lingüística, lo que ha traído como consecuencia que los estudios de comunidades epistémicas, los procesos de formulación y análisis de información sean casi nulos en estas herramientas (en su mayoría comerciales), lo que incide en la baja calidad de los servicios que prestan. Al olvidar a la Ciencia de la Información como disciplina totalizadora, que engloba los análisis de dominios y la descripción de mecanismos de búsqueda y recuperación de información, se ha soslayado un gran número de posibilidades metodológicas y pragmáticas sobre el tratamiento de la información.
- Las características de la hipertextualidad obliga a que los modelos de representación textual construidos a partir de paradigmas clásicos, se encaminen hacia modelizaciones multiesquemáticas, sustentadas en jerarquías heterogéneas, donde el resumen basado en elementos

semánticos se erija como un instrumento contentivo de estrategias de organización inteligentes, construidas bajo el estudio de complejos procesos heurísticos que trasciendan el plano fisicalista para insertase en un plano contextual.

- El modelo conceptual para la extracción de resúmenes de corpus textuales, desarrollados en el marco de esta investigación, se perfila como una solución metodológica acorde al problema científico planteado, pues facilita la generación de sistemas para procesamiento de textos y la confección de resúmenes a partir de documentos con una estructura retórica especifica. Además este modelo teórico conceptual y el conjunto de procedimientos descritos en él sirven de base para la creación de una herramienta, cuya puesta en práctica facilite la aplicación de la minería de textos en sistemas de bibliotecas.
- Los aportes del modelo al terreno del resumen automático residen en el desarrollo de técnicas de marcado para tratar los textos utilizando las capacidades del lenguaje XML y las propiedades de Python, el análisis de domino utilizando herramientas descritas en los estudios de discurso, tratadas en muchas investigaciones dentro de esta parcela de la Representación de la Información, pero escasamente modeladas para ser puestas en práctica, debido muchas veces, a la falta de solidez metodológica cuando son descritos los procesos. Por otra parte el uso de ontologías sobre las que se basan sistemas de agrupamiento de información y sistemas para desambiguar texto le confieren al modelo un valor socicognitivo, que se equipara con los nuevos tratamientos heurísticos demandados por la Ciencia de la Información para el proceso textual.
- El Análisis de Discurso y la lexicometría facilitaron la construcción de reglas heurísticas para que los agentes propuestos en el modelo puedan dotar de Inteligencia al acto de resumen, además con el estudio del discurso del dominio Ingeniería de Puertos y Costas, se declaran reglas léxicas, reglas de análisis y reglas de desambiguación, por lo que esta técnica no solo tributa información a los programadores para construir agentes, también facilita el estudio del corpus de modo que el modelo se estructure sobre las propia pragmática del domino.

- Se construyó una aplicación donde se pone de manifiesto todo el espectro teórico establecido en el modelo. Puertotext, es un software capaz de resumir, procesar y construir textos, que además permite la visualización, la navegación y el almacenamiento de información teniendo en cuenta los postulados de la web semántica , lo que permite el uso de una infraestructura general para la Representación del Conocimiento de modo que la información no sea propietaria de las aplicaciones que la utilizan. La Web semántica está orientada al intercambio de información entre humanos, al utilizarla para construir resúmenes, se facilitaría además el intercambio y la inferencia de nuevo conocimiento por parte de las propias máquinas y la modelización de nuevos procesos de extracción. Otro elemento de elevada condición cognitiva en la aplicación son los agentes, encargados, desde una perspectiva cognitiva, de representar las estrategias de conocimiento que asumen los integrantes del dominio para desarrollar sus textos.
- Múltiples y variadas son las técnicas que corroboran la calidad obtenida con la aplicación de este modelo. Los resúmenes que se obtienen con este sistema son de elevada calidad en comparación con los resultados reportados en otros modelos de extracción de textos. Para lograr estos resultados se apeló a las siguientes formas de evaluación: selección de términos, evaluación de ontologías, Rouge, evaluación heurística, de usabilidad y paneles de jueces, esto muestra un análisis del sistema desde que se acopian sus materias primas hasta que se obtiene un resultado pragmático, inclusive permitiendo la comparación a nivel de técnicas para demostrar fehacientemente la calidad del sistema y del modelo de forma integral.
- Siguiendo las concepciones metodológicas exigidas para este tipo de investigación se desarrolló una valoración del modelo por criterio de expertos, en la que fue corroborada la calidad del modelo y las posibilidades para que este se implemente en otros dominios de cocimiento siguiendo las mismas concepciones.


# Capítulo IX

## Trabajo Futuro

# Capítulo 9: Líneas Futuras de Investigación

En este apartado se van a enumerar los distintos desarrollos que a partir de ahora se van a llevar a cabo en el software PUERTOTEX y sobre los postulados teóricos de la investigación. Las líneas que se proponen son posibles campos de actuación a seguir una vez concluida la tesis, es importante aclarar que hay muchos de estos campos de actuación que ya se vienen desarrollando en la tesis y otros son meras propuestas para el futuro.

## *9.1.- Anotaciones Semánticas*

La necesidad de generar nuevas formas de anotación semántica en el texto son más que imprescindibles para el desarrollo del sistema. La importancia de la anotación de corpus reside en que es mayor la calidad que se alcanza en el software que se operan sobre corpus anotados. Hasta ahora en nuestra investigación solo se ha apelado al uso de anotaciones de cohesión en el texto, realizadas de forma semi-automática, pero el desarrollo de la aplicación demanda nuevas anotaciones y el desarrollo de un marcador discursivo lo suficientemente potente para marcar en el texto cada uno de los elementos semánticos.

### 9.1.1.- La anotación semántica: definición, tipos y técnicas

Una anotación es una descripción de elementos internos de un objeto, es la especificación de los atributos que corresponden a una entidad. Cuando nos referimos a objeto, estamos declarando elementos tales como: documentos, clases de una ontología, imágenes, etc.

Según Senso (Senso, 2009) la clasificación más extendida y aceptada por la comunidad es la que propuso Leech en 1997, la cual fue reformulada en 2004 (Leech, 1997, Leech, 2004), considerando seis tipos de anotación, a continuación se muestra las que necesita nuestro modelo:

- ***Part-of-speech (*POS):** que consiste en suministrar marcas morfológicas y de índole gramatical a una oración o a cada uno de los elementos que la componen. Según Senso (Senso, 2009) es un proceder donde se

identifica las partes de una oración, como son el nombre, verbo, artículo. En nuestro modelo este proceso debe ser mejor ya que solo se limita a declarar los verbos.

- **Lematización:** proceso que facilita añadir la procedencia de un lema en determinado corpus. Como bien afirma Senso (Senso, 2009) en inglés, la lematización puede considerarse redundante pero, en lenguas más flexivas como el español o el alemán, puede resultar de una gran utilidad para la extracción de información.
- **Anotación sintáctica:** Consiste en el suministro de identificadores de la sintaxis oracional a un corpus. Si bien en este modelo se apela a esta visión, es indiscutible que los elementos que se declaran en este apartado son insuficientes.
- **La anotación semántica**: consiste en añadir información acerca de la categoría semántica de las palabras.
- **La anotación del discurso**: Declaración de los elementos anafórico y catafóricos en el texto.
- **La anotación pragmática:** Este es el tipo de anotación aplicada al desarrollo de las reglas de discurso, si bien es el más desarrollado es el que más automatización demanda.

### 9.1.2. – Parsers Léxicos

Un parser o analizador sintáctico es una herramienta que facilita el análisis de las estructura de conjuntos de datos (Louden, 1997). En la literatura de la especialidad abundan las herramientas que se encargan de este proceso, sin embargo ninguna realiza las actividades que demanda el marcado de la estructura comunicativa. A continuación se declaran los métodos más utilizados para la construcción de parser según Louden y Aho (Louden, 1997, Aho et al., 1990):

- ***Métodos descendentes:*** Son procedimientos en los que se parte del un símbolo de inicialización en la gramática, que es posicionado en la en la raíz, lo que permite construir el árbol en forma descendente, es decir desde arriba hacia abajo hasta las hojas. Cuando se desarrolla este

método es vital elegir una derivación que genere a una concordancia con la cadena de entrada. Esta forma de construcción toma como centro el supuesto de *predice*, en el cual una derivación establece una concordancia en los símbolos de entrada (*predict/match*). Este tipo de análisis sintáctico descendente se realiza prefijando un recorrido delimitado de ante mano sobre el árbol.

- ***Métodos ascendentes:*** Este tipo de método se centra en la construcción de un árbol para el análisis sintáctico desde las hojas hasta la raíz. Cada hoja aloja a la cadena analizada, la cual debe ser reducida de forma idéntica al carácter de entrada de la gramática declarado en la raíz. Este proceder no es más que la técnica que facilita el desplazamiento en la cadena de entrada, para localizar una sub-cadena donde pueda aplicarse una reducción (*shift-reduce*). Como declaran Aho (Aho et al., 1990) el análisis sintáctico ascendente corresponde con un recorrido post-orden del árbol (primero reconocemos los hijos y luego mediante una reducción reconocemos el padre).
- ***Métodos direccionales****:* Compilan la cadena de entrada procesando cada símbolo por separado de izquierda a derecha.
- **Métodos no-direccionales:** Mediante estos métodos es posible acceder a cualquier segmento de la cadena de entrada para edificar el árbol. Estos métodos son muy costosos y tienen como premisa la necesidad de tener toda la cadena de componentes léxicos para implementarse.
- ***Métodos deterministas:*** A partir de la selección de un símbolo de la cadena de entrada es posible determinar una alternativa o derivación adecuada. No permite realizar retrocesos y solo determina una sola y única alternativa. Su coste operacional es lineal.
- ***Métodos no-deterministas: Obligan a determinar*** en cada paso de la construcción del árbol diferentes alternativas/derivaciones que propicien la identificación de la ruta adecuada de acuerdo con el coste operacional.

#### 9.1.2.1.- Herramienta de Análisis Sintáctico

Las herramientas que realizan análisis sintáctico son variadas, sin embargo sus prestaciones están centradas exclusivamente al trabajo con lematización, derivación y sintaxis. A continuación se explicitan las herramientas que se utilizan para estos análisis y los sitios donde se encuentran:

- **Thera:** Analizador sintáctico desarrollado por la universidad de Barcelona. Esta herramienta se encarga del análisis de oraciones y frases del español, el estudio morfológico de las palabras a partir de árboles, además posee herramientas muy potentes para flexionar términos, localizar nombres propios, localización de sinónimos y los significados de las palabras. Se puede consultar en http://clic.fil.ub.es/demo_sintactico/ 1. (Ver Figura 141)


Figura 141. Thera

- **Visual Interactive Syntax Learning:** Es un parser automático que se utiliza para el análisis sintáctico de textos en varios idiomas. Es facturado por Synddansk Universited. Tiene opciones para analizar oraciones, etiquetar corpus en varios idiomas y una amplia gama de herramientas de minería de texto. Permite visualizar en forma de grafo los resultados y exportarlos a otros formatos. http://visl.sdu.dk/visl/es/parsing/automatic/trees.php (Ver Figura 142)

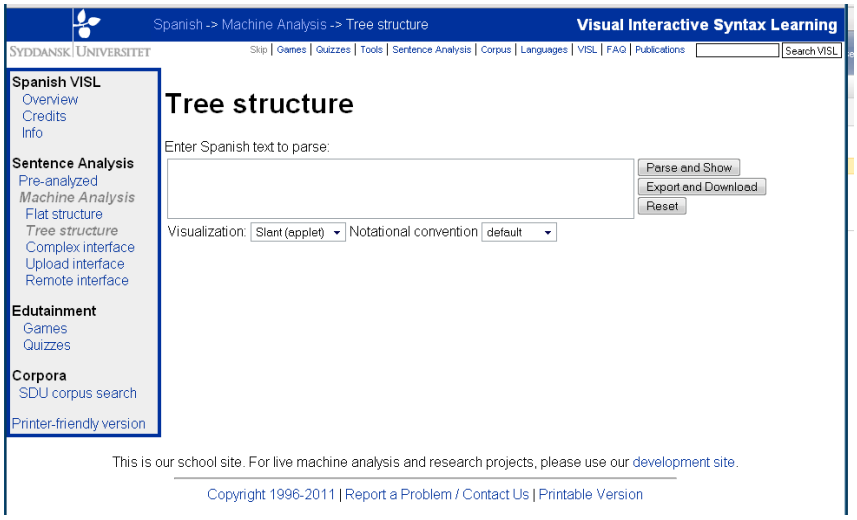



Figura 142. Visual Interactive Syntax Learning

EGD ha patrocinado un excelente desambiguador morfo-sintáctico para el español, desarrollado por el Grupo de Estructuras de Datos y lingüística Computacional de España. Entre las prestaciones que ofrece este software están sus capacidades la desambiguación, lematización, análisis sintáctico de oraciones, flexiones en vocablos, nombres propios y sinónimos. Se localiza en http://www.gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm (Ver Figura 143)


Figura 143. Desambiguador Morfosintáctico

- Otra de las aplicaciones de mayor nivel en el español es Stilus, de la corporación Dedalus. Esta herramienta provee a sus usuarios de innumerables facilidades para el análisis morfológico, conjugaciones verbales, juegos de caracteres, diccionarios y revisión de textos, identificando errores de posición y colación, además corrige algunos casos de coherencia, cohesión y anáfora pronominal. http://stilus.daedalus.es/demoIL.php? (Figura 144 )

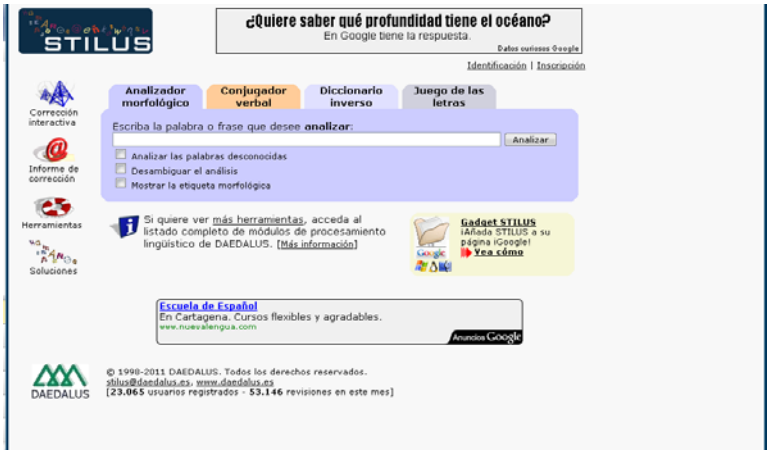



Figura 144. Stilus

ENGCG. Es un software desarrollado por Lingsoft y se encarga del análisis sintáctico de palabras del inglés, su única limitante reside en la poca capacidad de palabras que permite analizar, solo 100. http://www2.lingsoft.fi/cgi-bin/engcg/ (Figura 145)

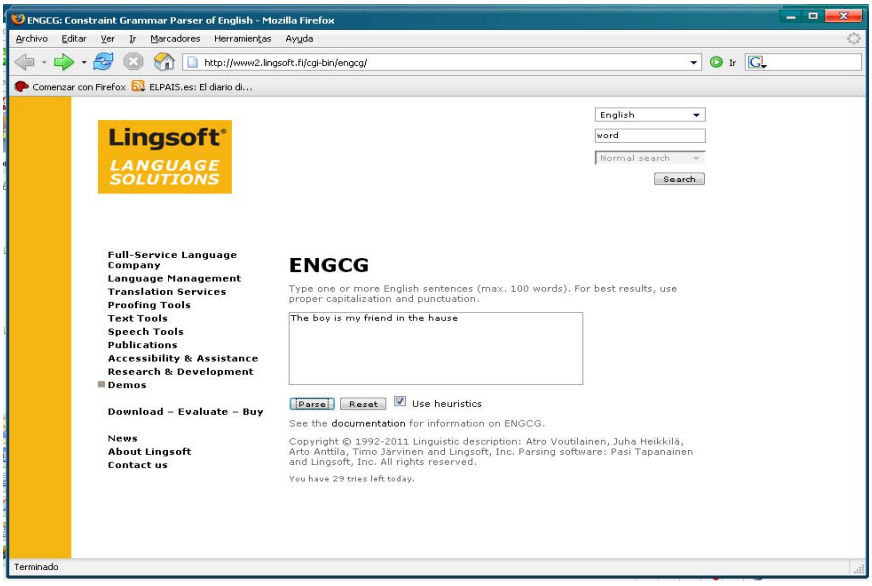


Figura 145. ENGCG

Lenguaje, es una firma dedicada a desarrollo de herramientas para el trabajo lingüístico y ha desarrollado un silabeador, que permite la división de palabras en idiomas español. http://www.lenguaje.com/herramientas/silabeador.php (Figura 146)

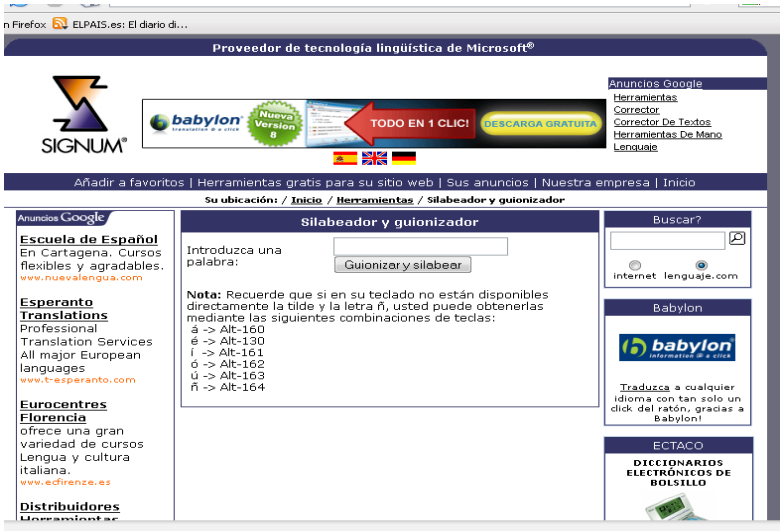



Figura 146. SIGNUM

Es evidente que no existe hoy en el español una herramienta que facilite la construcción de relaciones discursivas, por ello es necesario generar para nuevas versiones un parser que minorice el tiempo de marcado de las estructuras sintáctico comunicativas.

### *9.2.- Optimización de Ontologías*

Aunque ha quedado demostrado que la metodología utilizada para la implementación de la herramienta cumple con los estándares de desarrollo de ontologías, debido a sus facilidades para desarrollar el conocimiento en el dominio que se propone, es importante aclarar que se han presentado complejidades en el manejo de la ontología debido a su gran volumen, que unido a su más de 28 000 términos hacen que las consultas (a pesar de usar herramientas de potencia como el RDF Lib de Python) sean complejas y lentas. Una línea de investigación que se propone a partir de esto es la formulación de estrategias metodológicas para el desarrollo de ontologías. Mucho más que

buscar micro-ontologías unidas a una ontología, nuestra futura línea de investigación se sustentará en el aprovechamiento de la construcción de ontologías evaluadas y depuradas para luego generar micro ontologías y macro-ontologías, utilizando en los sistemas solo aquellos elementos que confluyen con el dominio de actuación, en donde la ontología se aplica.

Al igual que en Senso (Senso et al., 2007, 2009) se prevé usar mapas conceptuales y mapas mentales, ya que existen muchos procesos del pensamiento lógico asociado a los dominios que no se explicita en los mapas conceptuales y de esta forma se facilita la construcción de relaciones conceptuales con una carga de pensamiento lógico. El gran problema que ocasionan los mapas conceptuales es la escasa posibilidad de control terminológico y las insuficiencias normalizativas (no poseen normas de construcción). Es importante destacar que sus construcciones se hacen casi siempre en léxicos que no son compatibles con XML, entre ellos Java, etc. Sin embargo a pesar de sus deficiencias estas herramientas permiten modelar de forma precisa el contenido de un domino. Según (Moreiro González and García Martul, 2005) los mapas conceptuales permiten:

- Identificación de los conceptos importantes de un dominio
- Clasificación de los más generales a los más específicos
- Puesta en relación del conjunto.

Por su parte los mapas mentales son muy eficientes para mostrar relaciones de pensamiento, lo que permitiría que la ontología tuviese además un razonamiento declarado a partir de las estrategias cognitivas del dominio. Ambas formas de declarar el conocimiento pueden ser utilizadas como mecanismos para la construcción inicial de la ontología, teniendo en cuenta que la complejidad de estas puede ser tal, que puedan crearse problemas de estructuración. Si bien a partir de este punto el proceso de desarrollo de la ontología no será más fácil, es innegable la calidad que tendrá el mismo. Ninguna de las metodologías para la construcción que se conocen facilita tales propósitos (Anexo 27)

### *9.3.- Optimización de las ontologías para su consulta*

Tal como se declara en el capítulo 6 de esta tesis el sistema que se propone cuenta de una ontología y una base de datos para construir las relaciones que no sean de "tipo de". Esta base de datos se hizo en MySQL para escribir las relaciones que se necesitaban para hacer más explícita la búsqueda y las inferencias del sistema, debido a que el editor de ontologías Protégé no es capaz de generar de forma explícita las relaciones semánticas de la ontología. Lo primero que habrá que hacer será trabajar en una herramienta de construcción de ontologías que sea más clara en la declaración de las relaciones semánticas en las ontologías. Las otras que actúan sobre owl no son eficientes en este renglón y además son de pago. La solución iría en manejar la implementación de cada clase con identificadores temporales para distribuir los datos. Otra deficiencia que posee Protégé está en sus pocas capacidades para generar estructuras de tipo FOAF (Friend of a Friend), lo que obliga a desarrollar mejoras en su estructura para construir redes sociales a partir de grandes cantidades de datos. Según Senso (Senso, 2009) la parte que está directamente relacionada con la ontología es aquella que especifica las relaciones que existen entre los términos, que en el caso de esta aplicación son las siguientes:

- Afecta a
- Compuesto de (material)
- Contrario de
- Delimitado por
- Estudia
- Representa
- Se hace con
- Tiene función
- Tiene lugar en
- Ubicado en
- Tipo de
- Parte de
- Resultado de

- Atributo de
- Mide
- Conoce a
- Trabaja con
- Reside en
- Trabaja en

Esta ontología posee un 83 % de relaciones “tipo de”, pues es la que más se representa en la taxonomía, esto obliga a estudiar a fondo las relaciones y añadirles propiedades con otro editor de ontología para poder representar en un mismo documento RDF todos los elementos contentivos de la ontología (Clases, Subclases, Instancias y ejemplos).

La capa ontológica se sitúa sobre la de la base de datos de relaciones a modo de base cocimiento, detectando en el documento RDF todas las posibles relaciones asociadas a un término y reescribiendo automáticamente las relaciones como una cadena de caracteres para luego ser interpretadas por Python en el momento en que se realiza la consulta. En este caso se han desechado las posibilidades que ofrecen otras herramientas al uso como D2RQ que sirve de medio de comunicación con la ontología en varios proyectos internacionales.

La gran dificultad que hoy tenemos con la ontología reside en la extrema lentitud del sistema, debido a que su operatoria está basada en un sistema complejo de inferencia de conocimiento, con heurísticas de elevada complejidad. Al tener clases tan grandes cada búsqueda se realiza en un margen de tiempo de al menos 48 segundos como media.

Esta situación estriba en una línea de investigación. A continuación se observan algunas propuestas de solución al desarrollo de la optimización de la ontología para realizar consultas complejas en el dominio de ingeniería de Puertos y Costas:

- ***Computación distribuida*** (“Grid computing”). Es un proceso que obliga a la división de la información en segmentos pequeños para ser

procesados a la misma vez. Cada nodo de la ontología es almacenado en una máquina diferente de la red, la cual hará la consulta de más eficiente. Esta idea se ha utilizado en ontologías sencillas y grandes ontologías, con buenos resultados (Pernas and Dantas, 2005), (Cannataro et al., 2006), (Senso et al., 2007).

- ***Árboles binarios***. Utilizados en sistemas que toman decisiones en tiempo real, son estructuras de datos capaces de modelar procesos, esta técnica funciona recorriendo el árbol en dos sentidos. Como declara Senso (Senso, 2009) en las ontologías , la clave está en que el sistema sea capaz de escoger rápidamente la/s ramas del árbol (nuestra taxonomía) donde se encuentren las clases que están siendo empleadas en la consulta.

### *9.4. – Evaluación de Resúmenes automáticos*

La evaluación de resúmenes por vía automática solo ha sido descrita y representada en el análisis de bi-gramas a través de la herramienta Rouge (Lin, 2004), siempre realizando la comparación del resumen resultante de un modelo contra otro. Si embargo, no son suficientes los resultados que se logran con este tipo de valoración de un texto, pues a menudo aparece un texto con gran similitud en una métrica y escasa coherencia y legibilidad real (deficiencias en la pragmática). Son escasos los modelos de evaluación que incluyen también una dimensión cognitiva de la comprensión del texto (Pinto, 2001), por ello se necesita de un modelo de evaluación que integre de forma coherente las dimensiones cognitivas y estructurales del dominio en todas sus interacciones. En esta tesis se ha abierto una propuesta de evaluación que incluye todas las dimensiones que influyen sobre la calidad de un resumen por medios automáticos:

- Evaluación de Corpus
- Evaluación de Ontologías
- Evaluación por Rouge
- Evaluación por Jueces
- Test de Usabilidad de Sistema

### *9.5.- Resumen de Otros textos*

Ningún modelo particular de representación parece cumplir los requisitos de Meadow (1992, citado por (Chu, 2003): discriminar entre diferentes entidades, identificar entidades similares, permitir la descripción exacta de entidades y minimizar la ambigüedad en la representación.

Las disciplinas y tecnologías que emergen a mediados del siglo XX suponen espaldarazos para el cambio en los modelos de representación(Smith, 1993), por tanto el procesamiento de la información textual, ha de poner al texto y a los usuarios en un nuevo *status quo*, donde no basta solo con construir sistemas y modelos para formular la búsqueda distribuida, si no modelos que abarquen todo el fenómeno de la textualidad.

Internet no es un segmento espacial donde se construyan o se emitan conceptos únicos y estáticos para representar información, por ello la representación de la información no solo estará sujeta al texto verbal, si no a otros textos generados de la heterogénea realidad comunicacional; para lograr tal afirmación se necesita la contextualización de cada representación lingüística y el abordaje de nuevos diseños para los modelos de resumen, aunque sean temporales e imperfectos. Como bien dice Hernández (Hernández, 2007) *lo importante es ofrecer en el marco de la Documentación representaciones que expliciten y compartan, no que excluyan teorías o enfoques, que se inserten en la dinámica del razonamiento contextual y así reconozcan en la práctica los basamentos del enfoque sociocognitivo*.

Es incuestionable el papel preponderante de los contenidos mutimedios en los nuevos documentos que circulan en la red de redes. La variedad de software al uso para el tratamiento del resumen textual demandan un cambio más pegado a la realidad de los productos y sistemas que existen en la red debido a que los espacios de actuación se han ampliado y al usuario hay que garantizarle más que un texto resumido, un instrumento exponente que concrete la validez, la credibilidad, la fiabilidad, la congruencia entre su necesidad de información en un espacio de conocimiento cada vez más dinámico, que al decir de Hernández (Hernández, 2007) es un espacio con el que los usuarios están más

familiarizados por razones de educación y de cercanía generacional con la integración tecnológica.

Lograr representar esta realidad en nuevos resúmenes obliga redimensionar las fórmulas para que los sistemas asuman nuevos textos en el procesamiento de la información. La interacción entre nuevas formas de texto y la realidad de los nuevos documentos que circulan en Internet es la que hará que un nuevo tipo de resumen con características multimedio no se presente como un mediador descontextualizado, presionado por la avalancha hipertextual y por estrategias de representación clásicas.

La construcción de resúmenes de otras formas textuales conlleva a un análisis conceptual mucho más sofisticado que los que existen hoy en día, por ello se espera que la necesaria interacción entre estrategias mentales y dimensiones culturales sea un punto privilegiado en los nuevos métodos y formas de resumen.

Según Hernández (Hernández, 2007) y Pinto (Pinto et al., 2002) *no es suficiente que para construir resúmenes sobre documentos multimedia se conozca un modelo de ciclo de vida que integre la gestión del conocimiento en toda su totalidad solamente desde el punto de vista teórico, resta renovar la mirada epistémica, queda por aceptar una realidad documental interactiva e interdisciplinar, falta por hacer coincidir los procesos prácticos con la perspectiva compleja y con otras teorías comunicativas, para convertirlos en procesos cognitivamente ergonómicos.*

Para desarrollar estas visiones se necesita de lo siguiente:

- la flexibilidad en las concepciones modélicas para que los dominios estén debidamente representados a partir del procesos de aprendizaje continuo.
- una gramática textual mucho más compleja que se centre en los postulados de la cibersemiótica, apoyada en los significados y en las connotaciones, no en las clásicas denotaciones estructurales de los contenidos,

- desarrollo de metodologías y procesos de colaboración que favorezcan la comprensión de los modelos de representación del conocimiento,
- una representación ontológica e individual de los entes de conocimiento que se involucran en la Red, no solo desde el dominio tecnológico, sino también desde la praxis de diversas disciplinas como la documentación, la lingüística, la cibersemiótica, la psicología, etc.,
- una diálogo heterogéneo que permita reconocer la necesidad de la mirada de la representación hacia entes netamente proxémicos y recursivos, en una constante adecuación a las necesidades de los dominios y
- una formación acorde al cambio.

De acuerdo a estos cambios se proponen en esta tesis algunas líneas de trabajo para abordar nuevas formas de resumen de texto:

- las particularidades de los dominios de conocimiento,
- las nuevas competencias que demanda la formación profesional en las asignaturas que se imparten en Documentación y Ciencia de la Información
- la segmentación de imágenes por métodos algorítmicos, donde se encuentra las técnicas de Umbralización, Región Creciente, los Clasificadores, los métodos de Agrupamiento, la técnica de Campos de Markov, las Redes Neuronales, los Métodos Deformables, los Métodos guiados por plantillas y las Técnicas de ajuste al modelo.

### *9.6.- Trabajo con CMS*

Ante la falta de CMS semánticos, se está demandando el uso de otras herramientas para darle mejor operatividad al Sistema. Una mirada a los CMS que se insertan en el software libre y que pueden ser utilizados desde Cuba, nos obliga a elegir dos alternativas: Joomla y la Plataforma Drupal. Joomla posee buenos niveles de anotación pero su interacción con RDF es primitiva, sin embargo Drupal está comenzando a manejar varios módulos para el trabajo de RDF, entre los que se encuentra RDF CCk, el cual permite generar FOAF para los usuarios de páginas Web y otros módulos que son necesarios para la

implementación de ontología en RDF., los cuales poseen una semántica estructurada que se encuentra dimensionada con las necesidades de la actual Web Semántica. Drupal también posee múltiples aplicaciones para usarse con XML y facilita el acceso a bases de datos, cuestiones muy necesarias para el desarrollo futuro de Puertotex.


# Referencias Bibliográficas

***Referencias Bibliográficas***

AHO, A. V., SETHI, R. & ULLMAN, J. D. 1990. *Compiladores:principios, técnicas y herramientas,*.

CANNATARO, M., GUZZI, P. H., MAZZA, T., TRADIGO, G. & VELTRI, P. 2006. Managing ontologies for grid computing. *Multiagent and Grid Systems archive,* 2**,** 29-44.

CHU, H. 2003. *Information representation and retrieval in the digital age,* Santa Mónica, Amblin.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

LEECH, G. 1997. Introducing Corpus Annotation. *In:* GARSIDE, R., LEECH, G. & MCENERY, T. (eds.) *Corpus Annotantion.* London, New York: Longman.

LEECH, G. 2004. *RE: BT Oasis Corpus of Speech-Act Annotated Telephone Dialogues.* Type to (TO BE DISTRIBUTED BY HCRC EDINBURGH, W. T. P. O. B.

LIN, C. 2004. ROUGE: A Package for Automatic Evaluation of Summaries *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004.* Barcelona, España.

LOUDEN, K. C. 1997. *Compiler Construction: Principles and Practice*.

MOREIRO GONZÁLEZ, J. A. & GARCÍA MARTUL, D. 2005. La visualización de la información en revistas electrónicas mediante la concurrencia de herramientas hipertextuales, mapas conceptuales, topic maps y ontologías. *In:* INFORMAÇÃO, I. D. C. D., ed. VI CINFORM (Encontro Nacional de Ciência da Informação), 15 de junio, Salvador, Brasil.

PERNAS, A. M. & DANTAS, M. 2005. Grid Computing Environment Using Ontology Based Service. *Lecture Notes in Computer Science,* 3516.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez

PINTO, M., GARCÍA MARCO, F. & AGUSTÍN, M. D. C. 2002. *Indización y resumen de documentos digitales y multimedia: técnicas y procedimientos,* Madrid, Ediciones Trea.

SENSO, J. 2009. *Representación del conocimiento en la Ingeniería de Puertos y Costas.* Proyecto Investigador, Universidad de Granada.

SENSO, J. A., MAGAÑA, P. J., FABER-BENITEZ, P. & VILA, M. M. 2007. Metodología para la estructuración del conocimiento de una disciplina: el caso de PuertoTerm. *El profesional de la Información,* 16**,** 591-604.

SMITH, E. 1993. On the shoulders of giants: From Boole to Shannon to Taube: The origins and development of computerized information from the mid-19th century to the present. *Information Technology and Libraries,* 12**,** 217-226.


# Capítulo X

## Aportaciones de la Tesis

# Capítulo 10: Aportaciones de la Tesis

Introducción:

Desde que se formuló y aprobó el tema de investigación se han desarrollado diversas publicaciones en varias fuentes de información. Muchas de estas publicaciones fueron realizadas con profesionales del terreno de la Ingeniería Informática, el Control Automático y las Ciencias Biológicas. También es importante destacar el desarrollo de publicaciones con el director de la tesis y los Sres. Pedro Hípola y Luis Villén (Ver Figura 147).

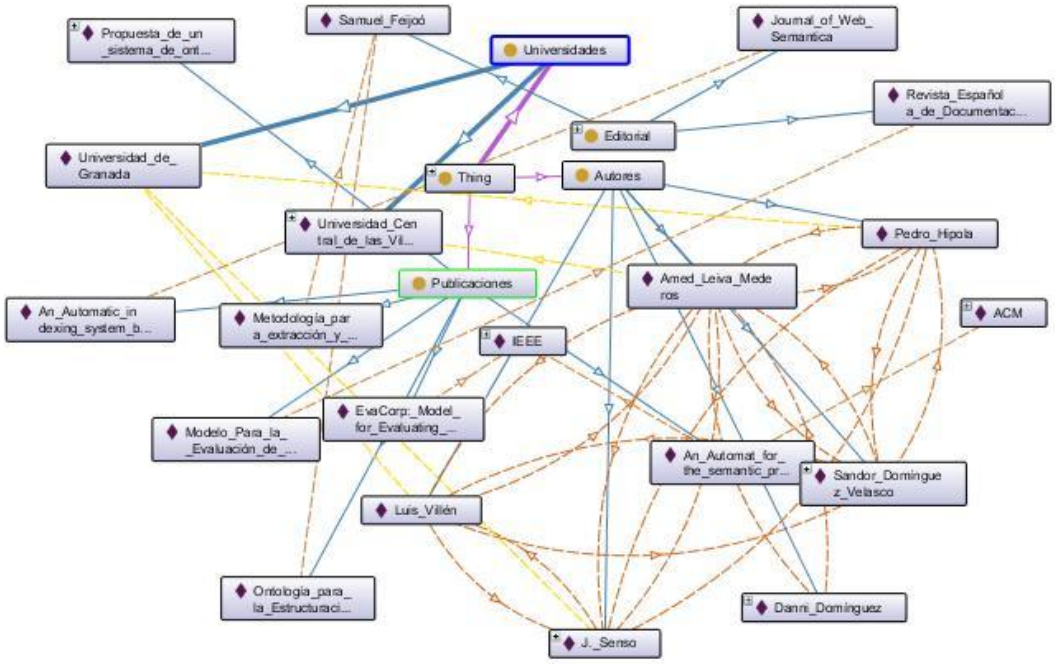



Figura 147. Relación de los Autores en los trabajos y aportaciones

A continuación se exponen de forma resumida los artículos presentados:

- **Senso, J., Leiva, A. 2008. *Metodología para extracción y desambiguación de textos científicos,* Santa Clara. Editorial Samuel Feijoo. ISBN: 978-959-250-427-1**. Es un libro donde los autores desarrollan un análisis de los métodos esenciales para la construcción de resúmenes, haciendo énfasis en las técnicas de

Minería de Texto y Minería de datos. Se analiza de forma crítica la poca explotación de los método de análisis de dominio en la construcción de extractos y se apuesta por la interdependencia de los modelos cognitivos en la solución a los problemas de estructura y cohesión en los resúmenes automáticos. Los autores combinan la Minería textual con herramientas ampliamente usadas en la Ciencia de la Computación para proponer una metodología particular que se basa en las bondades que han reportado diversos campos del saber a la Ciencia de lo Información. Finalmente se arriba a conclusiones y recomendaciones sobre el tema.

- **Leiva, A, Estévez, V. & Senso, J. 2008 Propuesta de un sistema de ontologías para la recuperación de recursos educativos. In Proceeding of Universidad 2008, Taller Internacional de Virtualización de la Enseñanza. ISBN 978-959—282-069-2**. Se exponen los principios esenciales de la Web Semántica y se propone un sistema ontológico capaz de servir de punto de acceso a diversos recursos educativos alojados en la red de la Universidad Central de las Villas, Cuba. Se arriba a conclusiones y Recomendaciones sobre el tema
- **Leiva, A, Estévez, V. & Senso, J.2008 Propuesta de un sistema de ontologías para la recuperación de recursos educativos. *Boletín Nueva Universidad*, No.1, V.1.** . Se exponen los principios esenciales de la Web Semántica y se propone un sistema ontológico capaz de servir de punto de acceso a diversos recursos educativos alojados en la red de la Universidad Central de las Villas, Cuba. Se arriba a conclusiones y Recomendaciones sobre el tema, además de presentarse una aplicación desarrollada en un servidor Apache que es capaz de brindar servicios a la comunidad Universitaria,
- **Leiva, A., Senso, J., Domínguez, S. & Hípola, P. 2009. *An Automat for the semantic processing of structured information.* In Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference On. Pissa,**

**Italia**. Se presenta un conjunto de agentes capaces de realizar el proceso de Indización aplicando algoritmos de agrupamiento. Se visualizan los resultados de los agrupamientos mediante una interfaz de Pajek y se arriba a conclusiones y recomendaciones sobre el tema propuesto.

- **Leiva, A, Domínguez, S., Domínguez, D., Senso, J. 2010 Ontología para la *Estructuración del conocimiento en Ecosistemas Marinos,* Santa Clara, Samuel Feijoó**. Texto que explicita una ontología que usa topónimos geográficos para localizar animales endémicos de la fauna cubana a través de anotaciones de segmentos de habitad.
- **Senso, J., Leiva, A., Domínguez, S. 2011 Modelo Para la Evaluación de Ontologías: aplicación en Ontosatcol. *REDOC. ISSN:* 0210-0614**. Después de un análisis crítico de la mayoría de los métodos de evaluación de ontologías los autores proponen un método de evaluación de ontologías centrado en las capacidades semánticas estructurales y lingüísticas de la ontologías, para la implementación del referido método usan como caso de estudio a Ontosatcol, una ontología para el dominio de Ingeniería de Puertos y Costas. Se arriba a Conclusiones y Recomendaciones sobre el tema.
- **Senso, J., Leiva, A., Domínguez, S. & Faber, P. 2011. Puertoterm: herramienta para la Recuperación de Recursos Sobre Medioambiente.** IX Evento Científico Bibliotecológico. Santa Clara, Cuba. Se declaran las capacidades de Puertoterm como herramienta lingüística para la Recuperación de la Información y las potencialidades de esta para la construcción de sistemas de búsqueda distribuida usando Top braid. Se arriba a conclusiones y recomendaciones sobre el sistema.

**Artículos en Evaluación por los Referees**

- **Senso, J., Leiva, A. & Domínguez, S. An Automatic indexing system based on semantic technologies. Journal of Web Semantic.** Propuesta de un sistema de agentes para lograr la indexación de artículos y documentos científicos. Se presentan

los agentes y un algoritmo que evalúa la calidad del agrupamiento de términos sobre una ontología, el cuál al ser comparado con otros algoritmos muestra la supremacía de cobra sobre la ontología

- **Senso, J. Leiva, A, Villén, L. & Domínguez, S. EvaCorp: Model for Evaluating Linguistic Corpora. Journal of Corpus Linguistics**. Se presenta una metodología para la selección de corpus textuales basada en procesos de agrupamiento y en análisis cualitativos. Finalmente se selecciona el corpus de mejor estructura para servir como modelo y el corpus de trabajo el que será sometido a normalización terminológica por reportar indicadores de evaluación muy bajos.


# Bibliografía

# Bibliografía

2003. *SPARQL query on the remote remote endpoint RDFLib / Redland* [Online]. Available: http://www.sparql-query-on-the-remote-remote-endpoint-rdflib-redland.htm [Accessed 26.abril 2011].

2004. *Introduction to using SPARQL to query an rdflib graph* [Online]. 2011. Available: http://www.IntroSparql.htm [Accessed 26. marzo].

ABADAL FALGUERAS, E. 2002. Elementos para la evaluación de interfaces de consulta de bases de datos web. *El Profesional de la Información,* 11**,** 349-360.

AGGARWAL, C. & YU, P. 2005. Online analysis of community evolution in data streams. *In Proceeding of SIAM* SIAM International Data Mining Conference.

AGGARWAL, P. & MUSTAFA, N. 2004 K-means projective clustering. *PODS.* Paris, France: ACM Press.

AHO, A. V., SETHI, R. & ULLMAN, J. D. 1990. *Compiladores:principios, técnicas y herramientas,*

AHONEN, H. 1988. Knowledge discovery in documents by extracting frequent word sequences. *Library trends,* 48 160-168.

AKAIKE, H. 1974. A new look at the statistical model identification. . *In:* IEEE (ed.).

ALLEN, J. 2005. *Natural Language Understanding,* London, The Benjamin Cummings Publishing.

ALLEN, M. 2005. *Estructuras de Datos y Algoritmos*, Addison-Wesley Iberoamericana.

ALONSO, J. 2002. *El Resumen,* Salamanca, Universidad de Salamanca, Facultad de Documentación y Traducción.

ALONSO, L. & FUENTES, M. 2002. Collaborating discourse for Text Summarisation *Proceedings of the Seventh ESSLLI Student Session. .* Trento.

ALONSO, M. 1958. Documentación. *Enciclopedia del idioma: diccionario histórico y moderno de la lengua español.* Madid: Aguilar.

ALVAREZ DE MON, I. 1999. *La Cohesión del texto .científico: un estudio contrastativo inglés-español.* Tesis Doctoral, Universidad Complutense de Madrid.

AMARO, L. 1977. *La redacción y el resumen*.

AMAT, N. 1988. *Documentación científica y nuevas tecnologías de la Información,* Madrid, Pirámide.

AMIGÓ, E. 2005. QARLA: A framework for the evaluation of text summarization systems. *In Proceeding of 43rd Annual Meeting of the Association for Computational Linguistic.* Michigan.

ANDERSEN, J. 2006. Knowledge organization: a sociohistorical analysys and critic. *Consulting Library Quarterly,* 76**,** 300-332.

ANKERST, M., BREUNIG, M., KRIEGEL, H. & SNADER, J. 1999. OPTICS: Ordering Points To Identify the Clustering Structure. *In Proceding of In International Conference on Management of Data Mining.* Philadelphia, PA, USA: ACM Press.

ANSI 1978. *American National Standard for writing abastracts,* NuevaYork, ANSI.

ARCO, L. 2007. *Corpus miner: herramienta para el etiquetado de grupos y la obtención de extractos.* MSc., Universidad Central de las Villas.

ARCO, L. 2008. *Agrupamiento basado en intermediación diferencial.* Tesis Doctoral, Universidad Central"Marta Abreu" de las Villas.

ARETOULAKI, M. 1996. *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis Of Texts For Their Automatic Summarization. .* Tesis doctoral University of Manchester.

ARETOULAKI, M. 1997. COSY-MATS: An Intelligent and Scalable Summarisation Shell. *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madrid.

ARROYO, M., AGUIRRE, J., BAVERA, F. & NORDIO, N. 2008. Jtlex Un Generador de Analizadores Léxicos Traductores. Argentina.

ASLAM, J., PELEKHOV, K. & RUS, D. 1998. Static and dynamic information organization withstar clusters. *In Proceeding of Conference of Information Knowledge Management.* Baltimore.

ATKINS, B. 1992. Tools for Computer-aided Corpus Lexicography: The Hector Project. *In:* KIEFER, F., KISS, G. & PAJZS, J. (ed.) *In Papers in Computational Lexicography. COMPLEX' 92.* Budapest: Linguistic Institute Hungarian Academy of Science.

AUSBEL, D. 1986. *Educational psychology: a cognitive view* New York, Holt, Rinehart & Winston.

AUSTIN, J. L. 1962. *How to do things with words. ,* Oxford: , Clarendon Press.

BACKER, F. & HUBERT, L. 1976. A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. *Journal of the American Statistical Association,* 71**,** 870-878.

BAEZA-YATES, R. & RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*, ACM Press ; Addison-Wesley.

BALDWIN, B., DONAWAY, R., HOVY, E., LIDDY, E., MANI, I., MARCU, D., MCKEOWN, K.MITTAL, WHITE, M. V., MOENS, M., RADEV, D., SPARCK-JONES, K., SUNDHEIM, B., TEUFEL, S. & WEISCHEDEL, R. 2000. An Evaluation Road Map for Summarization Research. *The Summarization Roadmap*.

BALDWIN, B. & MORTON, T. 1998. Co-reference-Based Summarization. *In:* FIRM, T., SUNDHEIM, B. (ed.) *In Proceeding of TIPSTER-SUMMAC Summarization Evaluation.* TIPSTER Text Phase III Workshop.

BANERJEE, A., KRUMPELMAN, C., BASU, S., MOONEY, R. & GHOSH, J. 2005. Model based overlapping clustering. *In Proceeding of International Conference on Knowledge Discovery and Data Mining (KDD).*

BAPTISTA, P., FERNÁNDEZ, C. & HERNÁNDEZ, R. 2005. *Metodología de la investigación,* La Habana, Pablo de la Torriente.

BARBUTI, R., GIACOBAZZI, R. & LEVI, G. 1993. A general framework for seman-tics-based bottom-up abstract interpretation of logic programs. *ACM Transactions on Programming Languages,* 15**,** 133-145.

BARNERS-LEE, T. & HENDLER, J. 2001. Scientific publishing on the 'semantic web. *Nature*.

BARNERS-LEE, T., HENDLER, J. & LASSILA, O. 2001. The semantic web. *Scientific American magazine,* 284**,** 34-43.

BARZILAY, R. 1988. Summarization evaluation methods: Experiments and analysis. *Proceeding of AAAI Intelligent Text Summarization Workshop.*

BARZILAY, R. & ELHADAD, M. 1999. *Using Lexical Chains for tex Summarization* [Online]. Negev. Available: http://acl.ldc.upenn.edu [Accessed 2008].

BATCHELOR, B. 1978. *Pattern Recognition: Idead in Practice,* Nueva York, Plenum Press.

BÉCUE, M. 1997. *Análisis Estadístico de Textos, cuarto seminario de capacitación de docentes,* PRESTA, Universidad de Concepción de Chile y Universidad libre de Bruxelles. Belgique. .

BÉLANGER, A. 2005. *Theory of summarization,* Canadá.

BELLO, R., ARCO, L. & ARTILES, M. 2006. New clustering validity measures based on roughset theory. *In:* FALCÓN, R. & BELLO, R. (eds.) *In Proceeding of International Symposium on Fuzzy and Rough Sets (ISFUROS'06).* Santa Clara, Cuba: Universidad Central"Marta Abreu" de las Villas, Facultad de Matemática Física y Computación.

BERGER, A. & MITTAL, V. 2000. A system for summarizing Web Pages. *Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval. .* Atenas.

BERGLER, S. 2004. *Multi-ERSS 2004, The CalC Laboratory* [Online]. Concordia: Departament of Computer Science. Concordia University. Available: http://www.nlpir.nist.gov [Accessed 26.feb. 2009].

BERNARAS, A., LARESGOITI, I. & CORERA, J. 1996. Building and Reusing Ontologies for ElectricalNetwork Applications. *In Proceeding of European Conference on Artificial Intelligence(ECAI96).*

BERRY, M. 2004. *Survey of Text mining: Clustering, Classification, and Retrieval,* New York, USA, Springer Verlag.

BEZDEK, J. & PAL, N. 1995. Cluster validation with generalized Dunn's indices. *In:* KASABOV, N., COGHILL, G. (ed.) *In Proceeding of 2nd International two-stream Conference on ANNES.* Piscataway, NJ: IEEE Press.

BIBER, D., CONRAD, S. & SPENSER, R. 1998. *Corpus linguistics investigating language structure and use,* Cambridge, Cambridge University Press.

BIRCKLEY, D. 2000. *The Friend of a Friend (FOAF) project* [Online]. Available: http://www.foaf-project.org/ [Accessed 26. mayo 20011].

BIRKLEY, T. 2000. *Models of ontolgy*, Willey.

BLACK, W. 1990. Knowledge-based abstracting. *Online Review,* 14**,** 227-240.

BLOOMFIELD, L. 1933. *Language,* Nueva York, Holt, Rinehart & Winston.

BOCK, H. 1985. On significance tests in cluster analysis. *J. Classification***,** 77-108.

BOEHM, B. W. 1981. *Software Engineering Economics.*

BOGURAEV, B. & KENNEDY, C. 1997. Salience-based content characterization of text documents. *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madrid. .

BOLELLI, L., ERTEKIN, S., ZHOU, D. & GILES, C. L. 2007. A clustering method for web data with multi-type interrelated components. *In Proceedings of 16th international conference on World Wide Web.* ACM Press.

BOLEY, D. 1988. Principal Direction Divisive Partitioning. *Data Mining and Knowledge Discovery,* 2**,** 325-344.

BOLÍVAR, A. 1998. *El discurso estilístico en el periodismo.* Tesis de Grado.

BONZI, S. 1991. Representation of concepts in text: a comparison of within document frequency, anaphora, and synonymy. *Canadian Journal of Information Science,* 16**,** 21-31.

BORDES, A., ERTEKIN, S., WESTON, J. & BOTTOU, L. 2005. Fast kernel classifiers with onlineand active learning. *Journal of Machine Learning Research***,** 1579-1619.

BORKO, H. & BERNIER, C. 1975. *Abstracting concepts and methods,* Nueva York, Academic Press.

BORKO, H. & CHATMAN, S. 1963. Criteria for aceptable abstracts: a survey of abstracters' instructions. *American Documentation*.

BOUILLON, P., CLAVEAU, V., FABRE, C. & SEBILLOTE, P. n.d. Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method.

BOWKER, L. & PEARSON, J. 2002. *Working with specialized language,* Londres y New York, Routledge.

BRANK, J., GROBELNIK, M. & MLADENIĆ, D. 2005. A survey of ontology evaluation techniques. *SIKDD 2005 at multiconference IS 2005.* Ljubljana, Slovenia.

BRASSARD, G. & BRATLEY, P. 1998. Fundamentos de Algoritmia. Prentice Hall.

BREWSTER, C., ALANI, H., DASMAHAPATRA, S. & WILKS, Y. 2004. Data Driven Ontology Evaluation. *Proceedings of the Language Resources and Evaluation Conference (LREC 2004).* Lisboa, Portugal.

BRICKLEY, D. & GUHA, R. 2000. Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation

BRIER, S. 2004. Cybersemiotics and the problems of the information-processing paradigm as a candidate for a unified Science of Information behind Library Information Science. *Library Trends,* 52**,** 629-657.

BRNADOW, R. & MITZE, Y. 2004. Automatic condensation of electronic publication by sentence selection. *Information Processign Management,* 43**,** 5,675-685.

BRON, J. & DAY, M. 1993. *Quality in abstracting or summarization,* New York, MC-Graw –Hill.

BROWN, A. L. & DAY, J. D. 1983. Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behavior,* 22**,** 1-14.

BRUN, M., SIMA, C., HUA, J., LOWEY, J., CARROLL, B., SUH, E. & DOUGHERTY, E. R. 2007. Model-based evaluation of clustering validation measures. *Pattern Recognition,* 40**,** 807-824.

BUZAN, T. 2003. *Mental Maps in knowledge,* Boston, Atterville.

CABADA, M. n.d. *[Automatización del servicio de resúmenes].* Tesis Doctoral, Universidad de la Habana.

CABRÉ, T., CODINA, L. & ESTOPÀ, R. (eds.) 2001. *Terminologia i documentació,* Barcelona: Institut Universitari de Lingüística Aplicada.

CALINSKI, R. & ARABAS, J. 1974. A dendrite method for cluster analysis. *Com.Statistics***,** 1-12.

CALLAN, J. 2005. Pasaje – level evidence in Document Retrieval. *In Proceedings of 17 International ACM SIGR Conference of Research of development in Information Retrieval.* ACM Press.

CANNATARO, M., GUZZI, P. H., MAZZA, T., TRADIGO, G. & VELTRI, P. 2006. Managing ontologies for grid computing. *Multiagent and Grid Systems archive,* 2**,** 29-44.

CAÑAS, A. 1995. Knowledge Construction and Sharing in Quorum. *In Proceeding of 7th World Conference on Artificial Intelligence in Education.* Washington, D.C.

CARLETTA, J. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics,* 22.

CARROLL, J., MINNEN, G. & BRISCOE, T. 1999. Corpus Annotation for Parser Evaluation. *EACL workshop on Linguistically Interpreted Corpora (LINC).* Bergen, Norway.

CASTILLO, E. & VÁZQUEZ, M. L. El rigor metodológico en la investigación cualitativa. *Colombia Médica,* 34.

CASTILLO, M. D. 2008. *Tendencias del resumen automático.* Universidad de la Habana.

CE OWDHURY, G. 1999. Template Mining for Information Extraction from digital documents. *Library Trends,* 48**,** 182-208.

CHACÓN, I. 2006. *La Mediación documental* [Online]. [Accessed 5.may. 2006].

CHAN, S. & TSOU, B. 1997. Discourse networkl: a framework for modelling textual structures. Pacific Association for Computational Linguistics.

CHANG-SHING, L., YEA-JUAN, C. & ZHI-WEI, J. 2003. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. *Expert Systems with Applications* 25**,** 431-447.

CHAUMIER, J. 1986. *Análisis y lenguajes documentales,* Barcelona, Mitre.

CHEN, P. & VERMA, R. 2006. A Query-based Medical Information Summarization System Using Ontology Knowledge. *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).* IEEE.

CHENG, D., KANNAN, R., VEMPALA, S. & WANG, G. 2006. A divide-and-merge methodology for clustering. ACM Trans. *Database Syst,* 31**,** 1499-1406.

CHENG, D., VEMPALA, S., KANNAN, R. & WANG, G. 2005. A divide-and-merge methodology for clustering. *In Proceeding of 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.* Baltimore, Maryland: ACM Press.

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

CLAUSÓ, A. 1996. Análisis documental formal: descripción y catalogación de monografías publicaciones periódicas. *In:* LÓPEZ YEPEZ, J. M. (ed.) *Manual de Documentación.* Madrid: Ediciones Pirámide.

CLEVELAND, D. & CLEVELAND, A. D. 2001. *Indexing and abstracting,* Englewood, Libraries Unlimited.

CLIMENT, S. 2006. *Sistemas de resumen automático de documentos* [Online]. Available: http//www.climet.blog._imp.htm [Accessed 3.junio 2006].

COL-VINENT, R. 1984. Las Operaciones de análisis documental. *En Ciencia documental: principios y Sistemas.* Barcelona: Editorial Mitre.

COMPANY, X. 2000. Inxight sofware:. Xerox enterprise company

CORNELIUS, I. 1997. Information and Interpretation. *In Proceedings of COLIS 2.* Copenhague: The Royal School of Libraiansahip.

CORRALES DEL CASTILLO, J. M. 2008. *Modelo de Servicio Semántico-Difuso de Difusión Selectiva de la Información para bibliotecas digitales.* PhD., Universidad de Granada, España.

CORTES, C., PREGIBON, D. & VOLINSKY, C. 2001. Communities of interest. *In Proceedings of 4th International Conference on Advances in Intelligent Data Analysis.*

COSERIU, E. 1992. *Competencia Lingüística,* Madrid, Gredos.

COYOTE, R. 2006. *Clasificación Automática de Textos considerando el Estilo de Redacción.* Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

CRAVEN, T. 1990. Use of words and phrases from full text in abstracts. *Journal of Information Science,* 16**,** 351-358.

CRAVEN, T. 2000. Abstracts Produced using Computer Assistance. *Journal of the American Society for Information Science,* 51**,** 745-756.

CREMMINS, E. 1985. *El arte de resumir,* Barcelona, Mitre.

CRESPO, T. 2007. *Respuestas a 16 preguntas sobre el empleo de expertos en la investigación pedagógica,* Perú, San Marcos.

CRONIN, B. 2008. The sociological turn in information science. *Journal of Information Science,* 34.

CUÉ, J. L. 1988. *Estadística,* La Habana, Pueblo y Educación.

CUNHA, I. D., FERNÁNDEZ, S. & VELÁZQUEZ MORALES, P. 2007. A New Hybrid Summarizer Based on Vector Space Model, Statistical Physics and Linguistics. *In:* GELBUKH, A. & KURI MORALES, F. (eds.) *MICAI 2007.* Berlín: Springer-Verlag Berlin Heidelberg.

D'CUNHA, I. 2006. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español.* Tesis Doctoral, Universidad Pompeu Fabra.

DAUDEN, M. 1982. *Redacción de documentos,* La Habana, Pueblo y Educación.

DAVE, R. 1996. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters,* 17**,** 613-623.

DENK, F. 2008. *Estructuras de Datos y Algoritmos* Wesley Iberoamericana.

DEODHAR, M. & GHOSH, J. 2007. A framework for simultaneous co-clustering and learning from complex data. *In Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Jose, California, USA: ACM Press.

DERBIN, B. n.d. *Gestión de Información* [Online]. Santa Clara, Cuba: Universidad Central "Marta Abreu" de las Villas. Available: http://intranet.cdict.uclv.edu.cu [Accessed marzo.26.2008 2008].

DÍAZ, J. 2003. Hipertexto periodístico: teoría y modelos. *In:* DÍAZ, J., SALAVERRÍA, R. (ed.) *Manual de redacción ciberperiodística.* Barcelona: Ariel.

DIK, S. 1978. *Functional Grammar,* Amsterdam, North Holland Publishing Company.

DIXON, M. 1997. *An Overview of Document Mining Technology* [Online]. Available:

http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dm.html [Accessed febrero.27. 2007].

DOMÍNGUEZ, S. 2009. Satcol 6 Herramienta Para El Minado De Corpus y Construcción De Índices Automáticos. *In:* BETA (ed.). Santa Clara: Universidad Central de las Villas, Departamento de Automática.

DOMÍNGUEZ, S. 2010. PROTEX. beta ed. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de automática.

DOMÍNGUEZ, S. 2011. Calculuscopora. beta ed. Santa Clara, Universidad Central"Marta Abreu" de las Villas: Departamento de Ingeniería Automática.

DOMÍNGUEZ, S. 2011. FOXCORP. Santa Clara, Cuba: Universidad Central de las Villas, Departamento de Ingeniería Automática.

DOMÍNGUEZ, S. 2011. Puertotex. Santa Clata: Universidad Central de las Villas, Departamento de Ingeniería en Control Automático.

DONAWAY, R., DRUMMEY, K. & MATHER, L. A. 2000. A Comparison of

Rankings Produced by Summarization Evaluation Measures. *In Proceeding of Workshop on Automatic Summarization.*

DOURISBOURE, Y., GERACI, F. & PELLEGRINI, M. 2007. Extraction and classification of densecommunities in the web. *In Proceedings of 16th international conference on World Wide Web.* . Banff, Alberta, Canada: ACM, Press.

DOWNEY, A., ELKNER, J. & MEYER, C. 2002. *Aprenda a Pensar como un Programador con Python,* Wellesley, Massachusetts, Green Tea Press.

DRÜSTELER, J. 2002. Information visualisation, what is it all about? *Inf@Vis! ,* 100.

DUNN, J. 1974. A fuzzy relative isodata process and its use in detecting compact well- separated clusters. *J. Cybernetics,* 3**,** 32-37.

DUNNING, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics,* 19.

EAGLES 1994. Corpus Typology: a framework for classification.Informe Interno No. 21. *In:* SINCLAIR, J. M. (ed.). Birminghan: Universidad de Birmingham, Corpus Linguistics Group.

EAGLES 1996. Preliminary Recommendations on Corpus Typology. Documento Eagles (Expert Advisory Group on Language Engineering) EAG-TCWG-CTYP/P.

EAGLES 1996. Text Corpora Working Group Reading Guide". Documento Eagles (Expert Advisory Group on Language Engineering) EAG-TCWG-FR-2.

EARL, L. 1970. Experiments in automatic extracting and indexing. *Information Storage and Retrieval,* 6**,** 313-323.

EDITORS, I. C. O. M. J. 2003. *Requieriments for manuscriips submitted to Biomedical Journals* [Online]. Vancouver. Available: http://www.cmaca/mwcuniform.htm [Accessed marz.10 2007].

EDITORS, I. C. O. M. J. 2003. *Uniform Requieriments for manuscriips submitted to Biomedical Journals* [Online]. Available: http://www.cmaca/mwcuniform.htm [Accessed febrero.5 2007].

EDMUNDSON, H. 1969. New Methods in automatic extracting. *Journal of the Asociation of Machinery,* 16**,** 264-285.

EDMUNDSON, H. 2006. *Methodology of abstracting Science,* Austin, Texas.

ENDRES-NIGGEMEYER, B. 1988. *Summarizing Information,* Hannover, Springer-Verlag.

ENDRES-NIGGEMEYER, B., MAIRE, E. Y SIGEL, A. 1995. How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. *Information Processing & Management,* 31**,** 631-674.

ENDRES-NIGGEMEYER, B. 2005. SimSum: an empirically founded simulation of summarizing *Information Processing and Management,* 36**,** 659-682.

EPTER, S. & KRISHNAMOORTHY, M. 1999. A multiple-resolution method for edge-centric dataclustering. *In Proceedings CIKM 1999 International*

*Conference on Information andKnowledge Management.* Kansas City, Misoury: ACM Press.

ESCOBAR, H. 2002. *Historia de las bibliotecas,* La Habana, Cuba, Sevigraf.

ESTADO, P. 2001. Recomendaciones para Obras Marítimas ROM 0.0 Procedimiento general y bases de cálculo en el proyecto de obras marítimas y portuarias, Madrid, Ministerio de Obras Públicas y Urbanismo.

ESTEBAN, M. A. 2006. Planificación, diseño y desarrollo de servicios de información digital. *In:* TRAMULLAS, J. & GARRIDO, P. (eds.) *Software libre para servicios de información digital.* Madrid: Pearson Prentice Hall.

ESTÉVEZ, V. 2005. *Puerto term: análisis perspectivo de un proyecto* [Online]. Santa Clara, Cuba: Universidad Central"Marta Abreu" de las Villas, CDICT. Available: http://intranet.cdict.uclv.edu.cu [Accessed 26.abril.2008 2008].

EVERMANN, J. E. & FANG, J. 2010. Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems,* 35**,** 391-403.

FABER, P. & MAIRAL-USÓN, R. 1999. *Constructing a lexicon of English verbs,* Berlín, Mouton de Gruyter.

FABER, P., MÁRQUEZ, C. & VEGA, M. 2005. Framing terminology: a process-oriented approach. *Meta,* 50**,** 189-213.

FABER, P., MONTERO MARTÍNEZ, S., CASTRO PRIETO, M. R., SENSO, J. A., PRIETO VELASCO, J. A., LEÓN ARAUZ, P., MÁRQUEZ LINARES, C. & VEGA EXPÓSITO, M. 2006. Process-oriented terminology management in the domain of coastal engineering. *Terminology* 2**,** 189-213.

FAHAD, M. & ABDUL QADIR, M. 2008. A Framework for Ontology Evaluation. *Supplementary Proceedings of the 16th International Conference on Conceptual Structures (ICCS'08).* Toulouse, France.

FALKOWSKI, T., BARTELHEIMER, J. & SPILIOPOULOU, M. 2006. Community Dynamics Mining. *In Proceedings of 14th European Conference on Information Systems.*

FELLBAUM, C. 1998. *WordNet - An Electronic Lexical Database*.

FENSEL, D. 2000. OIL in a nutshell. *In Proceedings of 12th European workshop knowledge acquisition, modelling and management.* New York: Springer-Verlang.

FENSEL, D. 2001. OIL: An ontology infrastructure for the semantic web. *IEEE Intelligent System* 16.

FERNÁNDEZ BREIS, J. T. 2003. *Un Entorno de Integración de Ontologías para el Desarrollo de Sistemas de Gestión de Conocimiento* Tesis Doctoral, Universidad de Murcia.

FERNÁNDEZ LÓPEZ, S. n.d. *SWAML, publicación de listas de correo en web semántica* Tesis de fin de Carrera, Universidad de Oviedo.

FERNÁNDEZ, O., TORAL, A. & MUÑOZ, R. n.d. Exploiting Lexical Measures and a Semantic LR to Tackle Textual Entailment in Italian.

FERNÁNDEZ, S., SANJUAN, E. & TORRES-MORENO, J. M. 2007. Énergie textuelle de mémoires associatives. *Actes de la conférence Traitement automatique des Langues Naturelles.* Tolouse.

FERNÁNDEZ-BREIS, J. T., EGAÑA ARANGUREN, M. & STEVENS, R. 2009. Quality evaluation framework for bio-ontologies. *In ICBO: International Conference on Biomedical Ontology, 2009* Buffalo, New York.

FERNÁNDEZ-LÓPEZ, M., GÓMEZ-PÉREZ, A. & JURISTO, N. 1997. Methontology: from ontological art towarads ontological engineering. *In Spring Symposium on Ontological. Engineering of AAAI, 1997.* California. : Stanford University.

FERRER, R. & SOLÉ, R. 2001. The small wolrd of human language. *Proc. R. Soc. Lond. B,* 268**,** 2261-2265.

FILLMORE, C. 1982. Frame Semantics. *In:* KOREA, T., L. S. O (ed.) *Linguistics in the morning calm.* Seoul, Hanshin.

FILLMORE, C. & ATKINS, S. 1998. FrameNet and Lexicographic Relevance. *Frist Internacional Conference on Languages Resources and Evaluation.* Granada.

FILLMORE, C., C., J. & PETRUCK, M. 2003. Background to Frame- Net. . *Journal of lexicography,* 16**,** 235-250.

FLORIA, A. 2000. *Evaluación Heurística* [Online]. Available: http://www.entrelinea.com/usabilidad/inspeccion/Heur.htm [Accessed 26.noviembre 2010].

FONER, L. N. 1996. What's an Agent, Anyway? Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. *In:* FRANKLIN, S. & GRAESSER, A. (eds.) *Third International Workshop on Agent Theories, Architectures and Languages.* Springer-Verlag.

FORD, K. 1991. ICONKAT: An Integrated Constructivist Knowledge Acquisition Tool. *Knowledge Acquisition Journal*.

FORD, K. 1996. Diagnosis and Explanation by a Nuclear Cardiology Expert System. *International Journal of Expert Systems*.

FORTUNATO, S., FREEMAN, L. & MENCZER, F. 2006. Scale-free network growth by ranking. *Physical Review Letters,* 96**,** 218701.

FOSKERT, W. 2008. Information on the text.

FRAKES, W. & BAEZA-YATES, R. 1992. *Information Retrieval :data Structure & Algorithms,* New York.

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.

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. .

FUKUMOTO, J. 2003. Text summarization based on itemized sentences and similar parts detection between documents. *In Proceedings of Third NTCIR Workshop.*

GADACHA, A. 2007. Assessment of metacognitive knowledge amongscience students, a case study of twobilingual and two NNS students. *System,* 35**,** 168-178.

GAHL, S. 1998. Automatic extraction of subcategorization frames for corpus-based dictionary making. *In Proceedings of Euralex'98*

GAIZAUSLAS – WILKS, J. 1988. Sistemas de trabajo con sumarización. Santa Clara, Cuba.

GANGEMI, A., CATENACCI, C., CIARAMITA, M. & LEHMANN, J. 2005. A Theoretical Framework for Ontology Evaluation and Validation.

GAO, B., LIU, T.-Y., ZHENG, X., CHENG, Q.-S. & MA, W.-Y. 2005. Consistent bipartite graphco-partitioning for star-structured high-order heterogeneous data co-clustering. *In Proceedings of 11thACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* Chicago, USA: ACM Press.

GARCÍA, A. 1984. Los lenguajes documentales. *In:* GARCÍA GUTIÉRREZ, A. (ed.) *Los Lenguajes documentales.* Barcelona: Mitre.

GARRIDO, M. 1996. Fundamentos del análisis documental. *In:* LÓPEZ YEPEZ, J. M. (ed.) *Manual de Información y documentación.* Madrid: Ediciones Pirámide.

GARVEY, S. & GRIFFIT, A. 2008. *[Algunas herramientas de los estudios de usuarios]* [Online]. Santa Clara: Unversidad Central de las Villas. Available: http://intranet.cdict.uclv.edu.cu [Accessed abril.25 2009].

GIBSON, D., KLEINBERG, J. & RAGHAVAN, P. 1998. Clustering categoricl data: an approach based on dynamical systems. *In Proceedings of 24th International Conference on Very Large Data Bases.* New York,USA: Morgan Kaufmann.

GIL, B. 1996. Lenguajes documentales. *In:* LÓPEZ YEPES, J. (ed.) *Manual de documentación.* Madrid: Ediciones Pirámide.

GIL, B. 2006. *Manual de lenguajes documentarios,* Madrid, Noesis.

GIL-GARCÍA, R., BADÍA-CONTELLES, J. & PONS-PORRATA, A. 2003. Extended Star clusteringalgorithm. *In Proceedings of CIARP.*

GIL-GARCÍA, R., BADÍA-CONTELLES, J. & PONS-PORRATA, A. 2006. A general framework for agglomerative hierarchical clustering algorithms.

*In Proceedings of 18th International Conference on Pattern Recognition (ICPR'06).*

GILL, K. & WHEDBEE, S. 2003. *Pragmatic in discurse*, Interamericana.

GIRONELLY, S. 1993. Contribución académica española al content analysis abstracting métodos, abstracts and information services. *Ciencia de la información,* 24**,** 121-124.

GOLDSTEIN, J. 1999. Summarizing Text Document: senetence Selection and Evaluation Metrics. *In Proceeding of SIGIR'99.*

GOLDSTEIN, J. 2000. *Creating and Evaluating Multi-document Sentence*.

GÓMEZ PÉREZ, A. 1994. Some Ideas and Examples to Evaluate Ontologies *In:* IEEE (ed.). Knowledge Systems Laboratory, Stanford University

GONZÁLEZ DUQUE, R. n.d. *Python para todos,* [s.l.], [s.n.].

GOODMAN, L. & KRUSKAL, W. 1954. Measures of associations for cross-validations. *J. Am. Stat. Assoc.,* 48**,** 732-764.

GOTLIEB, G. & KUMAR, S. 1968. Semantic clustering of index terms. *Journal of the ACM (JACM),* 15.

GOWER, J. & ROSS, G. 1969. Minimum spanning trees and single-linkage cluster analysis. *Applied Statistics,* 18**,** 54-64.

GRANOLLERS, T., LORÉS, J. & CAÑAS, J. J. 2005. *Diseño de sistemas interactivos centrados en el usuario,* Barcelona, UOC.

GRIMES, U. E. 1975. *The Thread of Discourse,* Mouton, The Hague.

GRUNINGER, M. F., M. 1995. The logic of enterprise modelling. *In:* SULLIVAN, J. B. D. O. (ed.) *Reenginering the Enterprise.* London Chapman & Hall.

GRUNNINGER, M. & FOX, M. S. 1995. Methodology for the Design and Evaluation of Ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95.* Montreal.

GUADARRAMA, P. 2009. *Dirección y asesoría de la investigación científica,* Bogotá, Editorial Magisterio.

GUARINO, N. & WELTY, C. 2002. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM,* 45**,** 61-65.

GUEREQUETA, R., VALLECILLO, A. 2000. *Técnicas de Diseño de Algoritmos,* Málaga, Servicio de Publicaciones de la Universidad de Málaga.

GUPTA, S. & STENT, A. J. n.d. Automatic Evaluation of Referring Expression Generation Using Corpora.

HALKIDI, M., BATISTAKIS, Y. & VAZIRGIANNIS, M. 2001. Clustering algorithms and validity measures. *13th International Conference on Scientific and Statistical Database Management.* IEEE Computer Society.

HALLIDAY, M. & HASAN, R. 1976. *Cohesion in English,* Essex, Longman.

HAN, J. & KAMBER, M. 2001. Data mining: concepts and techniques.

HAND, T. 1997. A proposal for task-based evaluation of Text Summarization Systems. *In Proceeding of ACL/EACL Wokshop on Intelligent Scalable Text Summarization,.*

HARMAN, D. & MARCU, D., ED. 2001. *Proceedings of the 1st Document Understanding Conference.* New Orleans, L.A.

HARTER, J. & BUSHA, L. 1990. *Metodología de la investigación en bibliotecología y Ciencia de la Información,* La Habana, Editorial Félix Varela.

HARTMANN, J., SPYNS, P., GIBOIN, A., MAYNARD, D., CUEL, R., SUÁREZ-FIGUEROA, M. C. & SURE, Y. 2005. *D1.2.3 Methods for ontology evaluation*, Knowledgeweb.

HATMANN, J., SPYNS, P., GIBOIN, A., MAYNARD, D., CUEL, R., SUÁREZ-FIGUEROA, M. C. & SURE, Y. 2005. D1.2.3 Methods for ontology evaluation, Knowledgeweb.

HAVENS, T. C., KELLER, J., POPESCU, M. & BEZDEK, J. C. 2008. Ontological Self-Organizing Maps for Cluster Visualization and Functional Summarization of Gene Products using Gene Ontology

Similarity Measures. *EEE International Conference on Fuzzy Systems (FUZZ 2008).*

HAYES-ROTH, B., PFLEGER, K., LALANDA, P., MORIGNOT, P. & BALABANOVIC, M. 1995. A Domain-Specific Software Architecture for Adaptive Intelligent Systems. *IEEE Trans. Sofware Eng* 21**,** 288-301.

HAYS, D. 1960. *Basic Principles and Technical Variations in Sentence Structure Determination,* Santa Mónica, RAND Corporation (Mathematical Division).

HAYS, D. 1964. *Dependence Theory: A Formalism and Some Observations Language 40*.

HENDLER, A. 2000. Sistems of text extraction.

HENLEY, J. 1993. Analytical abstract on CD-ROM. *Online &r CDROM Review,* 17**,** 285-290.

HENNIG, L., UMBRATH, W. & WETZKER, R. 2008. An Ontology-based Approach to Text Summarization. *EEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* Princenton: IEEE.

HERDAN, G. 1960. *Type-Token Mathematics: A Textbook of Mathematical Linguistics,* The Hague, The Netherlands, The Netherlands: Mouton & Co.

HERNÁNDEZ, A. 2006. indización y Resumen. La Habana: Universidad de la Habana.

HERNÁNDEZ, A. 2007. *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica.* Tesis Doctoral, Universidad de la Habana.

HERNÁNDEZ, A. 2007. *Paradigmas dominantes de la Representación de la Información y el Conocimiento.* Universidad de la Habana.

HERNÁNDEZ, E. 2006. *Algoritmo de Clustering basado en entropía para descubrir grupos en atributos de tipo mixto.* Instituto Politécnico Nacional.

HERNÁNDEZ, J., HERNÁNDEZ, R. & MEDINA, J. 2007. Compressed Arrays Algorithm for Frequent Patterns. *In:* MANSO, R. (ed.) *In Proceedings of SITIO 2007.* Santa Clara, Cuba: Editorial Samuel Feijoó.

HERNÁNDEZ, L. & MEDINA, J. 2007. Un Algoritmo para la Segmentación de Documentos por Tópicos. *In:* MANSO, R. (ed.) *In Proceedings of STIO 2007.* Snta Clara, Cuba: Editorial Samuel Feijoó.

HERRERA, R. 2007. *Formatos de Comunicación* [Online]. La Habana: Universidad de la Habana, Cuba. Available: http://fcom.uh.edu.cu [Accessed 16.julio 2009].

HERRERO-SOLANA, V. & RÍOS-GÓMEZ, M. 2006. Producción latinoamericana en biblioteconomía y documentación en el Social Science Citation Index *El profesional de la Información,* 15.

HINNEBURG, A. & KEIM, D. 1998. An efficient approach to clustering in large multimedia databases with noise. *In Proceedings of 4th International Conference on Knowledge Discovery and Data Mining.* New York, NY, USA: AAAI Press.

HJØRLAND, B. 2003. Fundamentals of knowledge organization. *Knowledge Organization,* 30**,** 87-11.

HJØRLAND, B. 2004. Domain analysis in information science. *Encyclopedia of Library and Information Science.* New York: Marcel Dekker.

HOCHBAUM, D. & SHMOYS, D. 1985. A best possible heuristic for the k-center problem. *Mathematics of Operations Research,* 10**,** 180-184.

HOEY, M. 1983. *On the Surface of Discoursa* London, George Alíen & Unwin.

HÖPPNER, F. 1999. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition West Sussex,* England, John Wiley & Sons

HOVY, E. & RADEV 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence,* 63**,** 341-361.

HSUN-HUI, H. & YAU-HWANG, K. 2007. Towards auto-contruction of domain ontology : an auto-constructed domain Conceptual lexicon and its application to extractive summarization

*Proceedings of the Sixth International Conference on Machine Learning and Cybernetics.* Hong Kong: IEEE.

HU, P., HE, T., JI, D. & WANG, M. 2004. A Study of Chinese Text Summarization Using Adaptive Clustering of Paragraphs. *Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04.* ACM.

HU, X. & WU, D. 2007. Data Mining and Predictive Modeling of Biomolecular Networkfrom Biomedical Literature Databases EEE/ACM Trans. *Comput. Biol.Bioinformatics,* 4**,** 251-253.

HUBERT, L. & SCHULTZ, J. 1976. Quadratic asignment as a general data-analysis strategy. *Br. J. Math. Stat. Psicol,* 29**,** 190-201.

HUFFMAN, D. A. 1952. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the I.R.E.*

HUNG, J. 2008. RETRACTION: A New WSD approach using word ontology and concept distribution. *Journal of Information Science* 31-48.

III, U. C. 1999. Resumir : reglas básicas *Revista electrónica de Información y Documentación*.

INGWERSEN, P. 1992. *Information retrieval interaction,* London, Taylor Graham.

IRAZAZÁBAL, A. 1996. Terminología y documentación. *Jornada panllatina de terminologia: perspectives i camps d'aplicación.*

ISO 1976. *Documentation. analyse pour les publications et ladocumentation,* Ginebra, ISO.

ISO 1985. *Guidelines for the establishment and development of multilingual thesauri,* Ginebra, International Organization for Standardization.

JAAN KAALEP, H. & VESKIS, K. n.d. Comparing Parallel Corpora and Evaluating their Quality.

JACKSON, P. 2001. *Multidocument text retrieval,* New York, Amblin.

JACKSON, P. & MOULINIER, I. (eds.) 2002. *Natural Language Processing for Online Applications* John Benjamins Publishing Company.

JACOBS, P. & RAU, L. 1993. Innovations in text interpretation. *Artificial Intelligence,* 63**,** 193-201.

JACOBSON, I., BOOCH, G. & RUMBAUGH, J. 2000. *El Proceso unificado de software,* Madrid, Addyson Wesley.

JAIN, A. & DUBES, R. 1988. *Algorithms for clustering data,* Englewood Cliffs, NJ, Prentice Hall College Div.

JAIN, A., MURTY, M. N. & FLYNN, P. J. 1999. Data clustering: a review. *ACM Computing Surveys,* 31**,** 264-276.

JAMES, A. & GUPTA, R. 2001. Information Retrieval for a summarizing. Center for Intelligent Information Retrieval Temporal Summaries of News Topics

JIANLIANG, X. & XIAOWEI, M. 2008. *International Conference on Advanced Language Processing and Web Information Technology.* IEEE.

JIMÉNEZ-HURTADO, C. 1994. El componente pragmático en el lexicón verbal del Español, alemán e inglés. Granada: Universidad de Granada, Departamento de Traducción e Interpretación.

JING, H. & MCKEOWN, K. 2000. Cut and paste based summarization. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics.* Seatle.

JIN-WOO, J., KYUNG-WOOK, P., JEONG HO, L., YOUNG SHIK, M., SUNG HAN, P. & DONG-HO, L. 2007. OLYVIA : Ontology-based Automatic Video Annotation and Summarization System using Semantic Inference Rules1 IEEE.

JOHNSON, F. C., GRIFFITHS, J. R. & HARTLEY, R. J. 2003. Task dimensions of user evaluations of information retrieval systems. *nformation Retrieval,* 18.

JONHSON, F. 1997. The apllication of linguistic processing to automatic abstract generation. *In:* JONES, K., WILLET, P. (ed.) *Readings in Information Retrieval.* San Francisco: Morgan Kaufman Publishers.

JONYER, I., COOK, D. & HOLDER, L. 2002. Graph-based hierarchical conceptual clustering. *Journal of Machine Learning Research,* 2**,** 19-43.

KALISKY, T., BRAUNSTEIN, L. A., SREENIVASAN, S., BULDYREV, S. V., HAVLIN, S. & STANLEY, H. E. 2006. Scale-free networks emerging from weighted random graphs. *Physical Review E,* 73**,** 025103.

KARYPIS, G., HAN, E. & KUMAR, V. 1999. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling. *IEEE Computer,* 32**,** 68-75.

KAUFMAN, L. & ROUSSEEUW, P. 1990. *Finding groups in data: an introduction to cluster analysis*, John Wiley.

KENT, A. 1955. Machine literature searching VIII. Operational Criteria for Designing Information Retrieval Systems. *American Documentation,* 6**,** 93-101.

KEPHART, J. & CHESS, D. 2003. The Vision of Autonomic Computing. *IEEE Computer***,** 41-50.

KERLINGER, F. 2006. *Investigación de comportamiento,* México, McGraw-Hill Interamericana.

KIM, D. & PARK, Y. 2001. A novel validity index for determination of the optimal number of clusters. *IEEE Trans. Inform. Syst., E84- D***,** 281-285.

KIM, M. & RAMAKRISHNA, R. 2005. New indices for cluster validity assessment. *Pattern Recognition Letters,* 26**,** 2353-2363.

KISSIL, J. 2006. Mapas conceptuales. CDICT, Universidad Central de las Villas.

KLEIN, H. 2001. *Text Analysis Softwares* [Online]. Available: http://www.intext.de.index.html [Accessed julio.23 2007].

KNAPP, A. 2002. *La experiencia del usuario,* Madrid, Anaya Multimedia.

KNIGHT, K. & MARCU, D. 2000. Statistics-based summarization – Step one: Sentence compression. *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence.* Texas.

KOEHN, P. n.d. Europarl: a multilingua corpus for evaluation of machine translation.

KRIEGEL, H. & PFEIFLE, M. 2005. Density-based clustering of uncertain data. *In Proceeding of 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* Chicago, Illionais, USA: ACM Press.

KRIPPENDORFF, K. 1991. *Metodología de análisis de contenido: teoría y práctica,* Barcelona, Paidós.

KROEKER, L. 2004. Seeing data: new methods for understanding information. *IEEE computer graphics and applications,* 24**,** 6-12.

KUHN, T. 1979. *The structure of scientific revolutions,,* Chicago, University of Chicago Press.

KUNCHEVA, L. & HADJITODOROV, S. 2004. Using diversity in cluster ensembles. *In Proceeding of IEEE SMC.* Netherlands.

KUPIEC, J. 2003. A trainable Document Summarizer. *In Proceeding of 18 th Annual International ACM SIGUIR Conference og rsearch Development in Information Retrieval.*

KUPIEC, J., PEDERSEN, J. O. & CHEN, F. 1995. A trainable document summarizer. *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95).* Seatle.

LABOV, W. & WALETSKY, J. 1967. Narrative analysis: Oral versions of personal experience. *Essays on the verbal and visual arts.* Seattle: University of Washington Press.

LAL, P. & RÜGER, S. 2002. Extract-based Summarization with Simplication. *Proceedings of the 2nd Document Understanding Conference at the 40th Meeting of the Association for Computational Linguistics.*

LANCASTER, F. 1990. *Indexing and abstracting in theory and practice*, University of Illinois, Graduate School of Library and Information Science.

LANCASTER, F. 1993. *Indización y resúmenes: teoría y prática*, Briquet de Lemos/libros.

LANCASTER, F. 1996. *El Control del vocabulario en la recuperación de la información.,* Valencia, Universidad de Valencia.

LANDAUER, T. & DUMAIS, S. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review,* 104**,** 211-240.

LANQUILLON, C. 2002. Enhancing Text Classification to Improve Information Filtering. *Künstliche Intelligenz***,** 37-38.

LAROCCA, J. & SANTOS, A. 2000. A trainable algorithm for summarizing news stories. *In Proceeding of 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD).*

LARSEN, B. & AONE, C. 1999. Fast and effective text mining using linear-time document clustering. *In Proceeding of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

LASSILA, O. & SWICK, R. 1999. Resource Description Framework (RDF) Model and Syntax Specification. *W3C Recommendation.*

LASSO, J. 1969. *Manual de Documentación,* Barcelona, Labor.

LEECH, G. 1997. Introducing Corpus Annotation. *In:* GARSIDE, R., LEECHA, G., MCENERY, T (ed.) *Corpus Annotantio.* Nueva York., London: Logman.

LEER, B. & COWY, I. 1996. *Sistemas de Sumarización*.

LEIVA, A. 2008. *Metodología para la extracción y desambiguación de textos científicos*

Tesis de Maestría, Universidad de la Habana.

LEIVA, A. 2009. *indización automática para alumnos de pregrado de la carrerra Bibliotecología y Ciencia de la Información* [Online]. Santa Clara:

Universidad Central de las Villas, CDICT. Available: http://sepad.fcie.uclv.edu.cu [Accessed 26.febrero 2009].

LEIVA, A., SENSO, J., DOMÍNGUEZ, S. & HÍPOLA, P. 2009. An Automat for the semantic processing of structured information. *In ISDA 9na International Conference of Desing of Software and Aplicactión.* Italia, Pissa: IEEE.

LENAT, D., GUHA, R. 1990. *Building large knowledge-based systems,* New York, Addison-Wesley Publising Company, Inc.

LEWANDOWSKA, B., (ED.) 1989. *Meaning and lexicography,* Amsterdam/ Philadelphia, John Benjamins.

LEWANDOWSKI, D. 2008. A three-year study on the freshness of web search engine databases. *Journal of Information Science***,** 817.

LEWIS, D. & RINGUETTE, M. 1994. A comparison of two learning algorithms for text classification. *In Proceeding of Third Annual Symposium on Document Analysis and Information Retrieval.* Las Vegas: Universidad de Nevada.

LEWIS, D. D. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *15th Ann Int'1 SIGIR '92.* Denmark.

LEZCANO, L. 2002. Modelos de tratamiento de textos y agrupamientos. Santa Clara, Cuba: Universidad Central"Marta Abreu" de las Villas.

LI, N. & MOTTA, E. 2010. Evaluations of User-Driven Ontology Summarization. *In:* CIMIANO, P. & PINTO, H. S. (eds.) *EEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* Springer-Verlag

LIN, C. 2004. Looking for a Few Good Metrics: Automatic Summarization Evaluation — How Many Samples Are Enough? *NTCIR Workshop 4,.* Tokyo, Japón.

LIN, C. 2004. ROUGE: A Package for Automatic Evaluation of Summaries *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004.* Barcelona, España.

LIN, C. & HOVY, E. 1998. Automatic Evaluation of summaries using n-gram co-ocurrence Statistic. *In Proceeding of HLTNAACL.* EE.UU.

LIN, C.-Y. & OCH, F. J. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics *42nd Annual Meeting of the Association for Computational Linguistics.* Barcelona, España.

LIN, C.-Y. & OCH, F. J. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation *20th International Conference on Computational Linguistics (COLING 2004).* Geneva, Switzerland.

LINARES, R. 2006. Bases Teóricas de la Ciencia de la Información: curso de la maetria de BCI (Bibliotecología y Ciencia de la Información (edición 6). La Habana: Universidad de la Habana.

LIU, Y., CAI, J., YIN, J. & HUANG, Z. 2006. An efficient clustering algorithm for small textdocuments. *In Proceeding of Seventh International Conference on Web-Age Information Management (WAIM 2006). .* IEEE Commnunications Society.

LONGACRE, R. E. 1983. Vertical threads of cohesion indiscourse. *In:* NEUBAUER, F. (ed.) *Coherence in Natural Language Texts'.* Hamburg,: Helmut Buske Verlag.

LÓPEZ, J. 1995. Propuesta de una guía para la elaboración de resúmenes más informativos. *ACIMED,* 3**,** 3-22.

LÓPEZ, J., COORD. 1996. *Manual de Información y Documentación,* Madrid, Pirámide.

LOPEZ YEPES, J. 2005. *Las tesis doctorales: producción, evaluación y defensa. .* Sevilla:, Fragua.

LÓPEZ-HUERTAS, M. 2008. Organización y representación del conocimiento: curso de doctorado. La Habana: Universidad de la Habana.

LÓPEZ-RODRÍGUEZ, C., TERCEDOR, M. & FABER, P. 2006. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. *Revista eSauld.com,* 2.

LOUDEN, K. C. 1997. *Compiler Construction: Principles and Practice*.

LOWEIN, B. 2004. *Hipermedial discurse,* Canadá.

LOZANO-TELLO, A. & GÓMEZ-PÉREZ, A. 2004. ONTOMETRIC: A method to choose the appropriate ontology. . *Journal of Database Management,* 15.

LUNH, H. 1958. The Automatic creation of Literature abstracts. *Journal of Research of Development***,** 159 – 165.

LUTZ, M. & ASCHER, D. 2003. *Python,* [s.l.], O'Reilly.

LYN, S. & HOVY, T. 2003. *Methods in summarization*, Pretice-Hall.

LYONS, J. 1977. *Semantics,* London, Cambridge University Press.

MAEDA, T. 1980. Automatic method for abstracting significant phrases in scientific or technical documents. *Information Processing & Management,* 16**,** 119-127.

MAEDCHE, A., MOTIK, B., STOJANOVIC, L., STUDER, R. & VOLZ, R. 2003. Ontologies for Enterprise Knowledge Management. *EEE Intelligent System***,** 26-34.

MAEDCHE, A. & STAAB, S. 2002. Measuring Similarity between Ontologies. *In: Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002, NCS/LNAI 2473.* Madrid, Spain.: Springer.

MAES, P. 1994. Social Interface Agents: Acquiring Competence by Learning from Users and Other Agents. *Working Notes of the AAAI Spring Syposium on Software Agents***,** 71-78.

MAGAN, J. 1996. Los Procesos técnicos. *Tratado básico de biblioteconomía.* Madrid: Editorial Complutense.

MAIGA, G. & DDEMBE, W. 2008. A Systems Approach to User Evaluation of Biomedical Ontologies *First International Workshop on Ontologies in Interactive Systems.* IEEE.

MANCHÓN, E. 2002. *Evaluación por criterios o heurística* [Online]. Available: http://www.ainda.info/evaluacion_heuristica.html [Accessed 1.enero 2011].

MANI, I. & BLOERDON, E. 1998. Machine learning of Generic and User-focused. *In Proceeding of the Fifteenth National Conference on Artificial Inteligence.* ACM Press.

MANI, I. & BLOERDON, E. 1998. Multi-document Summarization by Graph Search and Matching. *In Proceedings of AAAI.* ACM Press.

MANI, I., GATES, B. & BLOERDORN, E. 1999. Improving Summaries by Revising Them. *In Proceeding of the 37th Annual Meeting of the ACL.*

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

MANIEZ, J. 1990. La Evolución de los lenguajes documentales (traducción)

MANN, W. & THOMPSON, S. 1990. Rhetorical structure theory: a theory of text organization.

MANN, W. C. & THOMPSON, S. A. 1986. Assertions from Discourse Structure. *In:* CONFERENCE, H. L. T. (ed.) *Workshop on Strategies Computing Natural Language.* Marina del Rey, California.

MANNIG, C., RAGHAVAN, P. & SCHÜTZE, H. 2008. *An Introduction to Information Retrieval,* Cambridge, Cambridge University Press.

MANNIG, C. & SHÜTZE, H. 2000. *Foundations of Statistical Natural Language Processing,* Cambridge, MA, MIT Press.

MAÑA, M. 2003. *Generación automática de resúmenes de texto para el acceso a la información.* Tesis doctoral, Universidad de Vigo.

MARCOS, M. C. & GÓMEZ, M. 2006. Idoneidad de las interfaces de léxicos y terminologías en la web : Glat: Aspects méthodologiques pour l'élaboration de lexiques unilingues et de multilingues. *Bertinoro***,** 17-20.

MARCU, D. 1997. The Rhetorical Parsing of Natural Language Texts. *In Proceeding of 35th Annual Meeting of the Association for Computational Linguistics, (ACL'97/EACL'97).* Madrid: ACM

MARCU, D. 1999. Discourse trees are good indicators of importance in text. *In:* MANI, I., MAYBURY, M. (ed.) *Advances in Automatic Text Summarization.* MIT Press.

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

MARIE, M. 2007. Summarizing court decisions. *Information Processing and Management,* 43**,** 1748–1764.

MARIMÓN CARRAZANA, J. A. 2005. *Aproximación al estudio del modelo como resultado científico,* Santa Clara, CENTRO DE ESTUDIOS PEDAGOGICOS.

MARINELLI, D. 2002. Los Sistemas para el desarrollo de algoritmos.

MARTÍN MINGORANCE, L. 1995. Lexical logic and structural semantics: methodological underpinnings in the structuring of a lexical database for natural language processing. *In:* HOINKES, L. (ed.) *Panorama der Lexikalishen Semantik.* Tubinga: Gunter Narr.

MARTÍN MINGORANCE, L. n.d. Functional Grammar and Lexamatics. *In:* TOMASZCZYK, J., LEWANDOWSKA, B. (ed.) *Meaning and Lexicography.* Amsterdam/Philadephia: John Benjamins.

MARTINELLI, D. A. n.d. *Identificación de hábitos de uso de sitios web utilizando redes neuronales.* Tesis de Grado.

MATEO, P., GONZÁLEZ, J. C., VILLENA, J. & MARTÍNEZ, J. L. 2003. Un sistema para resumen automático de textos en castellano. *Procesamiento del Lenguaje Natural,* 31**,** 29-36.

MATHIS, B., RUSH, J. & YOUNG, C. 1973. Improvement of automatic abstracts by the use of structural analysis. *JASIS,* 24**,** 101-109.

MATLCLAF, W. 2006. *A Biblography Research in text Summarization* [Online]. Available: http://www.si.umich.edu/~radev/summarization/large-bib.doc [Accessed 23.marzo.2006 2006].

MAULIK, U. & BANDYOPADHYAY, S. 2002. Performance evaluation of some clustering algorithms and validity indices. *Pattern Anal Mach Intell,,* 24**,** 1650-1654.

MAYBURY, M. 1995. Generating Summaries from event Data. *Information Processing and Management* 31**,** 735-751.

MAZAL, A. & GARCÍA, I. 2006. *Introducción a la programación con Python,* [s.l.], Departamento de Lenguajes y Sistemas Informáticos Universidad Jaume.

MCENERY, A. & WILSON, A. 1996. *Corpus Linguistics,* Edimburgo, Edimburg University Press.

MCKEOWN, K. & RADEV, D. 1995. Generating summaries of multiple news articles. *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95).* Seattle.

MEANS, S. 2007. Representaciones del VSM.

MEDINA, J., HECHEVARRÍA, J. & GONZALEZ, B. 2007. Experimentación de algoritmos. *In:* MANSO, R. (ed.) *Sitio 2007.* Santa Clara: CDICT.

MEDINA, J. & PÉREZ, A. 2007. ACONS: a new algorithm for clustering. *CIARP.*

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

MEL'CUK, I. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. *In:* AGEL, V., EICHINNGER, L., EROMS, H., HELLWIG, P., HERRINGER, H. J. & LOBIN, H. (eds.) *Dependency and Valency :an International Handbook of Contemporary Research.* Berlín - Nueva York: W. de Gruyter.

MENESES PLACERES, G. 2010. *ALFINEV: Propuesta de un modelo para la evaluaciónde la alfabetización informacional en la Educación Superior en Cuba* PhD., Universidad de la Habana.

MIHALCEA, R., CORLEY, C. & STRAPPARAVA, C. n.d. Corpus-based and Knowledge-based Measures of Text Semantic Similarity

MILLER, G. 1993. *Introduction to WordNet: An On-line Lexical Database.* [Online]. Available: http://elies.rediris.es/elies9/2-4-2.htm [Accessed 26.mayo 2008].

MILLER, G. 2001. The W3C semantic web activity. *International Semantic Web Workshop.*

MILLIGAN, G. & COOPER, M. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* 50.

MING-CHENG, T., WEN-YANG, L. & RONG, J. 2008. Incremental maintenance of generalized association rules under taxonomy evolution *Journal of Information Science* 34.

MIZARRO, S. & TASCO, L. n.d. *Recuperación de la información* [Online]. Available: en: http// ftp.cdict.uclv.edu.cu [Accessed abril.28 2008].

MIZARRO, S. & TASO, C. 2004. Ephemeral and persistent personalization in adaptive information acces to Scholarly publications on the Web. *Segunda Conferencia internacional de Hipermedia Adaptativa.* Málaga.

MLADENIC, D. & GROBELNIK, M. 1998. Feature selection for classification based on text hierarchy. *Conference on Automatic Learning and Discovery (CONALD-98).*

MOLONES, J. 2004. Introduction of semantic information in Information Science.

MONTALVO, C. 2006. *Evaluacion de la Seleccion, Traduccion y Pesado de los Rasgos para la Mejora del Clustering Multilingüe* [Online]. Available: http://www.CMPI.2006.pdf [Accessed 25.mayo 2008].

MONTAÑO RAMÍREZ, A. n.d. *Python*.

MONTEJO-RÁEZ, A. 2005. Algoritmos de alta densidad.

MONTES Y GÓMEZ, M. & VILLASEÑOR PINEDA, L. 1996. Desambiguación delSentido de las Palabras (Word Sense Disambiguation).

MOREIRO, J. 1989. El resumen científico en el contexto de la teoría de la documentación. *Documentación de las Ciencias de la Información* 12**,** 147-170.

MOREIRO, J. 1996. La Técnica del resumen científico. *In:* LÓPEZ YEPEZ, J. (ed.) *Manual de Información y documentación.* Madrid: Ediciones Pirámide.

MOREIRO, J. 2004. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural,* Madrid, Ediciones Trea.

MOREIRO, J. 2006. *El resumen científico en el contexto de la teoría de la documentación. Texto y descripción sustancial* [Online]. Madrid. Available: http://www.ucm.es [Accessed 26.octubre 2006].

MORENO, L. 1996. Recensión, reseña y revisión en el marco de las actividades documentales : precisiones documentales. *Documentación de las Ciencias de la información,* 19**,** 211-234.

MORRIS, A., KASPER, G. & ADAMS, D. 1992. The efecets and limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research,* 3.

MÜLLER, R., SPILIOPOULOU, M. & LENZ, H. 2005. The influence of incentives and culture on knowledge sharing. *In proceeding of 38th Hawaii International Conference on System Sciences (HICSS-38 2005).* Big Island, Hawaii, USA IEEE Computer Society.

NAMBA, L. 1999. *Information Retrieval* Appleton.

NANBA, H. & OKUMURA, M. P. 2000. Producing More Readable Extracts by Revising Them. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000).* Saarbrücken.

NEDOBITY, W. 1982. The relevance of terminologies for automatic abstracting. *Journal of Information Science,* 4**,** 161-165.

NEGROPONTE, N. 1995. *El mundo digital,* Barcelona, Ediciones B.

NELSON., T. 1992. *Literary Machines,* Sausalito, Mindful Press.

NICHOLSON, S. 1997. Indexing and Abstracting on the World Wide Web: an examination of six web databases. *Information Technology and Libraries***,** 1-14.

NIELSEN, J. 1994. Heuristic evaluation. *In:* NIELSEN, J. & MACK, R. (eds.) *Usability Inspection Methods.* New York, NY.: John Wiley & Sons,.

NIELSEN, J. 2002. *How to Conduct a Heuristic Evaluation* [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_evaluation.html [Accessed 26. enero 2011].

NIELSEN, J. 2002. *Ten Usability Heuristics* [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_list.html [Accessed 21.enero 2011].

NIGAM, K., MCCALLUM, A., THRUN, S. & MITCHELL, T. 2000. Text classification from la belled and unlabeled documents using EM. *Machine Learning,* 38**,** 103-115.

NING, H. & SHIHAN, D. 2006. Structure-Based Ontology Evaluation. *In: International Conference on e-business on Engineering.* Computer Society, IEEE.

NIU, Z., JI, D. & TAN, C. 2004. Document clustering based on cluster validation. *CIKM'04 Thirteenth ACM International Conference on Information and Knowledge Management.* ACM Press.

NOMOTO, J. & MATSUMOTO, S. 2001. A New Approach to Unsupervised Text Summarization. *Proceedigs of SIGIR.*

NORMALIZACIÓN, C. C. E. D. (ed.) 1983. *Resúmenes y Anotaciones (NC 39-12)* La Habana: CEN.

NOVAK, J. 1991. Concept maps and vee diagrams: two metacognitive tools to facilitate meaningful learning. *Instructional Science,* 19**,** 1-25.

NOVAK, J. & GOWIN, D. 1985. *Learning How to Learn,* New York, Cambridge University Press.

NÚÑEZ, I. 2005. *AMIGA.* Tesis Doctoral, Universidad de la Habana.

OBESO, A. 2001. *Algunas cuestiones relativas al resumen automático y el agrupamiento,* México, D.F., UNAM.

OBRST, L., WERNER, C., INDERJEET, M., RAY, S. & SMITH, B. 2007. The Evaluation of Ontologies: Toward Improved Semantic Interoperability. *In:* BAKER, C. J. O. & CHEUNG, K.-H. (eds.) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences.* Springer.

OLDAKOWSKI, C., BIZER, C. & WESTPHAL, D. 2005. RAP: RDF API for PHP. *Workshop on Scripting for the Semantic Web.* Heraklion.

ONCINIZ-MARTÍNEZ, J. L. 2009. Towards a Corpus-Based Analysis of Anglicisms in Spanish: A Case Study. *International Journal of English Studies*, 115-132.

ONO, K., SUMITA, K. & MIIKE, S. 1994. Abstract generation based on rhetorical structure extraction. *Proceedings of the International Conference on Computational Linguistics.* Kyoto.

ORASAN, C. & BOROVETS. 2007. Pronominal anaphora resolution for text summarisation. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007). .*

ORDOÑEZ, C. & OMIECINSKI, E. 2002. FREM: fast and robust EM clustering for large datasets. *CIKM '02 (eleventh international conference on Information).*

ORLANDIC, R., LAI, Y. & YEE, W. 2005. Clustering high-dimensional data using an efficient and effective data space reduction. *In 14th ACM International Conference on Information and Knowledge Management.* Bremen, Germany: ACM Press.

OTERO, A. 2007. *Algoritmos estructurales basados en lógica borrosa para la identificación y caracterización de apneas e hipoapneas* [Online]. Available: http://www.otero.blog [Accessed 27.septiembre 2007].

OU, S. 2008. Design and development of a concept-based multi-document summarization system for research abstracs. *Journal of Information Science.,* 34.

PAICE, C. & JONES, P. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. *16th International Conference on Research.* ACM Press.

PAICE, D. 1988. The automatic generation of abstracts of technical papers. Lancester: Lancester University Press.

PAICE, D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management,* 26, 71-186.

PAICE, D. 1994. Automatic abstracting. *Enciclopedia of Library and Information Science.* New York: Marcel Dekker.

PAICE, D. 2002. *Construcción de resúmenes de único texto*.

PANECA, F. 2009. *La Prensa remediana del siglo XIX: reflejo de la cultura popular tradicional de la localidad.* Tesis de Grado, Universidad Central "Marta Abreu" de las Villas.

PARSONS, K., MCCORMAC, A. & BUTAVICIUS, M. 2009. *Human Dimensions of Corpora Comparison: an Analysis of Kilgarriff's (2001) Approach,* Melbourne, Command, Control, Communications and Intelligence Division Defence Science and Technology Organisation.

PASLARU BONTAS, E. & MOCHOL, M. 2005. A Cost Model for Ontology Engineering. *In:* BERLIN, F. U. (ed.).

PAST, K. 2008. The Giant: an agent-based Approach to Knowledge Construction & Sharing. *Eleventh Florida Artificial Intelligence Research Symposium.* Santa isabel island, Florida, USA.

PERALTA, M., LEIVA, A., ESTÉVEZ, V. & RUÍZ, M. 2006. *Retos y tendencias de la representación de la información y el conocimiento,* Santa Clara, Cuba, Editorial Samuel Feijoó.

PÉRES HERNÁNDEZ, M. C. 2002. *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento.* Málaga, Universidad de Málaga.

PÉREZ, A. & PAGOLA, J. 2007. Documentación de Gstar y ACONS. La Habana: CENATAV.

PÉREZ, J. & MEDINA, J. 2007. A Generalized Star Algorithm for Clustering. *In:* MANSO, R. (ed.) *SITIO 2007.* Santa Clara, Cuba: Editorial Samuel Feijoó.

PÉREZ-ÁLVAREZ, J. 1998. *Introducción a la información y documentación científica,* Madrid, Alhambra/Universidad.

PERNAS, A. M. & DANTAS, M. 2005. Grid Computing Environment Using Ontology Based Service. *Lecture Notes in Computer Science,* 3516.

PETER, G. & ROSEMANN, M. 2000. Integrated Process modeling: an ontological evaluation *Information System,* 25**,** 73-87.

PINEDA, L. A., CASTELLANOS, H., CUÉTARAB, J., GALESCU, L., JUÁREZ, J., LLISTERRID, J., PÉREZA, P. & VILLASEÑORE, L. n.d. The Corpus DIMEx100: Transcription and Evaluation.

PINK, A. 1997. Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science,* 48**,** 382-394.

PINTO AVEDAÑO, D. E. 2008. *On Clustering and Evaluation of Narrow Domain Short-Text Corpora.* PHD, Universidad Polit´ecnica de Valencia, Departamento de Sistemas Informáticos y Computación, Reconocimiento de Formas e Inteligencia Artificial.

PINTO, M. 1996. Análisis documental de contenido. *In:* LÓPEZ YEPEZ, J. (ed.) *Manual de Información y Documentación.* Madrid: Pirámide.

PINTO, M. 2001. *El resumen documental: principios y métodos,* Madrid, Fundación Germán Sánchez Ruipérez

PINTO, M. 2003. Engineering the Production of Meta-Information: The Abstracting Concern. *Journal of Information Science* 405.

PINTO, M. 2004. Interdisciplinary Approaches to the Concept and Practice ofWritten Text Documentary Content Analysis. *Journal of Documentation,* 50**,** 405-418.

PINTO, M. n.d. Representación de la información y el Procesamiento del Lenguaje Natural.

PINTO, M. & CORDÓN, J.-M. E. 1999. *Técnicas documentales aplicadas a la traducción,* Madrid, Síntesis.

PINTO, M., GARCÍA MARCO, F. & AGUSTÍN, M. D. C. 2002. *Indización y resumen de documentos digitales y multimedia: técnicas y procedimientos,* Madrid, Ediciones Trea.

PIRRÒ, G. & TALIA, D. 2008. LOM: a linguistic ontology matcher based on information retrieval. *Journal of Information Science* 34.

PONS, A. 2006. Una panorámica de la construcción de extractos de un texto. *Revista Cubana de Ciencias Informáticas,* 2**,** 55-67.

POPESCU, M., KELLER, J. M., MITCHELL, T., BEZDEK, J. C. & 2004. Functional Summarization of Gene Product Clusters Using Gene Ontology Similarity Measures. IEEE.

PORTER, M. 1980. An algorithm for suffix stripping. *Program,* 14**,** 130-137.

PSICOLOGÍA, U. F. D. 2004. *Desarrollo y comunicación de documentos hipermediales.,* Santa Clara, Cuba, Editorial Samuel Feijoó.

QADIR, M. & NOSHAIRWAN, W. 2007. Las ontologías: advertencias para la omisión del Conocimiento en disyuntas Las ontologías. *Segunda Conferencia Internacional sobre Internet y sus aplicaciones Web y Servicios (ICIW07), 2007.*: . IEEE.

QADIR, M. & NOSHAIRWAN, W. 2007. Warnings for Disjoint Knowledge Omission in Ontologies. *Second International Conference on internet and Web Applications and Services (ICIW07).* IEEE.

QUIAN, Y., ZHANG, G. & ZHANG, K. 2004. FA\ÇADE: a fast and effective approach to thediscovery of dense clusters in noisy spatial data. *SIGMOD '04:2004 ACM SIGMOD International Conference on Management of Data.* Paris, France: ACM Press.

RADEV, D. 1999. *Language Reuse and Regeneration: Generating Natural Language*

*Summaries from Multiple On-Line Sources.* Thesis Doctoral, Columbia University.

RADEV, D. 2001. Experiment in single and multidocument summarization using MEAD. *1st Understanding Conference.*

RADEV, D., BLAIR-GOLDENSOHN, S., ZHANG, Z. & S., R. R. 2001. Interactive, Domain-Independent Identification and Summarization of Topically Related News Article. *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries.* Londres.

RADEV, D., HOVY, E. & MCKEOWN, K., (EDS.) 2002. *Computational Linguistics (4)Special Issue on Summarization*, The MIT Press.

RADICCHI, F., CASTELLANO, C., CECCONI, F., LORETO, V. & PARISI, D. 2004. Defining and identifying communities *Networks. PNAS.* USA: National Academy of Science.

RAILEANU, D., BUITELAAR, P., VINTAR, S. & BAY, J. 1999. Evaluation Corpora for Sense Disambiguation in the Medical Domain.

RAJAN, K., RAMALINGAM, V., GANESAN, M., PALANIVEL, S. & PALANIAPPAN, B. 2009. Automatic classification of Tamil documents using vector space model and artificial neural network. *Expert Systems with Applications,* 36**,** 10914–10918.

RAMÍREZ, Z. 2007. *El Análisis del dominio en la organización y representación del conocimiento.* Diploma de Estudios Avanzados, Universidad de Granada.

RAMÍREZ, Z. & MONTES DE OCA, A. 2004. *Redes de información,* La Habana, Félix Varela.

RAMOS, E., NÚÑEZ, H. & CASAÑAS, R. 2009. Esquemas para evaluar ontologías únicas para un dominio de conocimiento. *Revista Venezolana de Información, Tecnología y Cococimiento,* 6**,** 57-71.

RAU, L. 1987. Knowledge Organization and Acces in a Conceptual Information System. *Information Processing and Management,* 23**,** 419-428.

REICHHERZER, T. 1998. *Models of summarizations and algorithm* [Online]. Available: http://www.ihmc.us [Accessed 23.febrero 2007].

RETCKEN, J. 2008. *Models of vocabulary* Manchester, Amblin.

RICHARDSON, C. 1999. Spatial Knowledge Acquisition from maps and form navigation in real and virtual environments. *Memory and Cognition,* 24**,** pp., 741-750.

RICO, C. 1994. *Aproximación estadístico - algebraica al problema de la resolución de la anáfora en el discurso.* Tesis Doctoral, Universidad de Alicante.

RITTBERGER, M. 1997. Measuring quality in the production of Databases. *Journal of Information Science,* 23.

ROBERT, P. & MALAKA, R. 2004. A Task-Based Approach for Ontology Evaluation. *ECAI Workshop on Ontology Learning and Population.* Valencia, Spain: European Media Laboratory GmbH.

ROBERTS, A., GAIZAUSKAS, R., HEPPLE, M., DEMETRIOU, G., GUO, Y., ROBERTS, I. & SETZER, A. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics,* 42**,** 950–966.

ROBINSON, J. 1970. A Dependency Based Transformational Grammar. *Actes du Xme Congrès international des linguistes 2.* Bucarets.

ROCHE, F., ÁLVAREZ, L. & ARJONA, L. n.d. Caracterización de Páginas Web para su Clasificación Automática. Santa Clara, Cuba: UCLV.

ROQUE, M. 2008. Mapas conceptuales. Santa Clara, Cuba: UCLV.

ROSELL, M., KANN, V. & LITTON, J. 2004. Comparing comparisons: document clustering evaluation using two manual classifications. *IIn Proceeding of CON 2004 International Conference on Natural Language Processing* Hyderabad, India.

ROSMAN, C. 1996. *La Sociedad Digital y su perspectiva sociológica,* Santa Clara, Editorial Samuel Feijoo.

ROUSSINOV, D. & CHEN, H. 1999. Document clustering for electronic meetings: an experimental comparison of two techniques. *Decision Support Systems,* 27**,** 67-79.

RUIZ-SHULCLOPER, J., ALBA-CABRERA, E. & SÁNCHEZ-DÍAZ, G. 2000. DGLC a density-basedglobal logical combinatorial clustering algorithm for large mixed incomplete data. *Symposium. IGARSS* IEEE

RUSH, J., SALVADOR, R. & ZAMORA, A. 1971. Automatic abastracting and indexing. II : production of indicative abstract. *Journal of the ASIS,* 22**,** 260-275.

RUSSELL, S. J. & NORVIG, P. 2003. *Artificial Intelligence : A Modern Approach*, Prentice Hall/Pearson Education.

SABOU, M., LÓPEZ, V., MOTTA, E. & UREN, V. 2006. Ontology Selection:Ontology Evaluation on the Real Semantic Web. *WWW2006.* Edimburgo.

SAGGION, H. & LAPALME, G. 2002. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics* 28**,** 497-526.

SAHAMI, M. 1998. *Using Machine Learning to Improve Information Access.* PhD. Thesis, Stanford University. .

SAHAMI, M., DUMAIS, D., HECKERMAN, D. & HOVITZ, E. 1988. A Bayessian approach to filtering junk a-mail.

SALES, D. 2006. *La biblioteca de Babel: documentarse para traducir,* Granada, Comares.

SALES-SALVADOR, D. 2006. *Documentación aplicada a la traducción: presente y futuro de una disciplina,* Gijón, Trea.

SALTON, G. 1996. On the application of syntactic methodologies in automatic text analysis. *Information Processing and Management,* 26**,** 73-92.

SALTON, G., ALLAN, J. 1997. Automatic Análisys Theme Generation, and Summarization of machina-readable Texts. *In:* SPARK, K., WILLET, P. (ed.) *Reading in Information Ridieval.* San Francisco: Morgan Kauffman.

SALTON, G. 1997. Automatic text structuring and summarization. *Information Processing and Management,* 33**,** 193-208.

SALTON, G. & BUCKLEY, C. 1988. Term weighting approaches. *Automatic text Information Processing and Management,* 24**,** 513-523.

SALTON, G. & MCGILLM, M. 1983. *Introduction to modern information retrieval.,* Nueva York:, McGraw-Hill.

SALTON, G., WONG, A. & YANG, C. S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,,* 18**,** 613-620.

SAN SEGUNDO, R. 1996. *Sistemas de Organización del Conocimiento,* Madrid, Pirámide.

SÁNCHEZ, A., SARMENTO, T., CANTOS, P. & SIMÓN, J. 1995. *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones,* Madrid, SGEL.

SANDERSON, M. 2005. Currated used directed summarization from existin tolls. *International Conference of Information and Knowledge Management.*

SANDIG, L. & SELTING, B. 2003. *Los retos del discurso espistémico,* México, Interamericana.

SATRIANO, C. & MOSCOLONI, N. 2000. Importancia del análisis textual como herramienta para el análisis del Discurso : aplicación en una investigación acerca de los abandonos del tratamiento en pacientes drogodependientes. *Cinta de Moebio***,** 1-23.

SCHAAL, M., MÜLLER, R., BRUNZEL, M. & SPILIOPOULOU, M. 2005. RELFIN - Topic discovery for ontology enhancement and annnotation. *In:* GÓMEZ-PÉREZ, A. & EUZENAT, J. (eds.) *In Proceeding of The Semantic Web: Research and Applications, Second European Semantic Web Symposium (ESWC 2005).* Crete, Greece: Springer, Heraklion.

SCHÜTZE, H. 1992. Dimensions of Meaning. *Proceedings of Supercomputing '92,.* Los Alamitos, California: IEEE Computer Society Press.

SCHWARTZ, G. 1978. Estimation the dimension of a model. *Ann Statu,* 6**,** pp. 461-464.

SEARLE, J. 1965. What is a speech act *In:* BLACK, M. (ed.) *Philosophy in Americ.* Londres: George Allen.

SEARLE, J. 1969. *Speech acts. An essay in the Philosophy of Language,* Cambridge, Cambridge University Press.

SEARLE, J. 1975. Indirect speech acts. *In:* COLE, P. & MORGAN, J. (eds.) *Syntax and Semantics 3. Speech Acts.* New York: Academic Press.

SENSO, J. 2008. *Descripción e intercambio en la web semántica* [Online]. Granada: Universidad de Granada. Available: http://documentacion.ugr.es [Accessed 28. septiembre 2010].

SENSO, J. 2009. *Representación del conocimiento en la Ingeniería de Puertos y Costas.* Proyecto Investigador, Universidad de Granada.

SENSO, J. & EÍTO, R. 2004. Minería textual. *El profesional de la información,* 13**,** 11-27.

SENSO, J. & LEIVA, A. 2008. *Metamodelo para la extracción y desambiguación de textos científicos,* Santa Clara, Cuba, Universidad Central "Marta Abreu" de las Villas, Editorial Samuel Feijoó.

SENSO, J., LEIVA, A. & DOMÍNGUEZ, S. 2011. Modelo para la evaluación de ontologías. Aplicación en Onto-Satcol. *Revista Española de Documentación Científica*.

SENSO, J. A., MAGAÑA, P. J., FABER-BENITEZ, P. & VILA, M. M. 2007. Metodología para la estructuración del conocimiento de una disciplina: el caso de PuertoTerm. *El profesional de la Información,* 16**,** 591-604.

SETKIN, M. 2006. *Summarization,* Hanover, Atertile.

SHAMS, R. & ELSAYED, A. 2008. A Corpus-based Evaluation of Lexical Components of a Domainspecific Text to Knowledge Mapping Prototype. *11th International Conference on Computer and Information Technology (ICCIT 2008).* Khulna, Bangladesh.

SHAMS, R., ELSAYED, A. & MAH- ZEREEN AKTER, Q. 2010. A Corpus-based Evaluation of a Domain-specific Text to Knowledge Mapping Prototype. *JOURNAL OF COMPUTERS,,* 5.

SHANNON, C. 1948. A mathematical theory of communications. *The Bell System Technical Journal of Artificial Intelligence Research,* 27**,** 379-423, 623-656.

SHEIKHOLESLAMI, G., CHATTERJEE, S. & ZHANG, A. 2000. WaveCluster: a wavelet-basedclustering approach for spatial data in very large databases. *The VLDB Journal,* 8**,** 289-304.

SIEGEL, S. & CASTELLAN, N. 1988. *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill.

SILBER, H. G. & MCCOY, K. F. 2000. Efficient Text Summarization Using Lexical Chains. *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000). .* Nueva York.

SILBERSCHATZ, A. & TUZHILIN, A. 1996. what makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Engineering,* 8.

SILVA, O. 2004. El análisis del discurso según Van Dijk y los estudios de la comunicación. *Razón y Palabra*

SMITH, D. C., CYPHER, A. & SPOHRER, J. 1994. Programming Agents without a Programming Language. *Communications of the ACM,* 37**,** 55-67.

SMITH, M., WELTY, C. & MCGUINNESS, D. 2004. OWL web ontology language guide. W3C Recommendation.

SMITH, T. F. & WATERMAN, M. S. 1981. Identication of common molecular subsequences. *Journal of Molecular Biology,* 147**,** 195-197.

SOLTEN, T. (2005). Sistema semiológico de la hipernmedia.

SPARCK-JONES, K. 2001. Factorial summary evaluation». *Proceedings of the Workshop on Text Summarization del ACM SIGIR Conference 2001.* Nueva Orleans. .

SPARK, K. 1988. Automatic summarising- factors an directions. *In:* MANI, I., MAYBURY, M. (ed.) *Advances in autliomatic text summarisation.* Cambridge,MA: MI I' Press.

SPILIOPOULOU, M., SCHAAL, M., MÜLLER, R. M. & BRUNZEL, M. 2005. Evaluation of Ontology Enhancement Tools. *In:* ACKERMANN, M. (ed.) *In Proceeding of Semantics, Web and Mining, Joint International Workshops.* Porto, Portugal: Springer.

SPINK, A. 1997. Interaction in information retrieval: selection and effectiveness of search terms. *Journal of the American Society for Information Science***,** 34 -39.

SPYNS, P. 2005. EvaLexon: Assessing triples mined from texts. Bruselas.

STAAB, S. 2004. Why Evaluate Ontology Technologies? Because It Works! *IEEE INTELLIGENT SYSTEMS*.

STAAB, S., SCHNUR, H., SUDER, R. & SURE, Y. 2001. Knowledge Processes and Ontologies. *IEEE Intelligent Systems,* 16**,** 28-39.

STACY, H. 2007. Task-based evaluation of text summarization using Relevance Prediction. *Information Processing and Management* 43**,** 1482–1499.

STEIN, G., BAGGA, A. & WISE, G. B. 2000. Multi-document summarization: Methodologies and evaluations. *TALN.*

STEINBACH, M., KARYPIS, G. & KUMAR, V. 2000. A comparison of document clustering techniques. *In Proceeding of KDD Workshop on Text Mining. Washington.* ACM Press.

STEPHENSON, K. & ZELEN, M. 1989. Rethinking centrality: methods and examples. *Social Networks***,** 1-37.

STEVENSON , L., RAVIN, Y. & LEACOCK, C. (eds.) 2000. *Large Vocabulary Word Sense Disambiguation Polysemy. Theoretical and Computational Approach,,* Oxford,: Oxford University Press.

STEVENSON, M. & WILKS, Y. 2000. Large Vocabulary Word Sense Disambiguation. *In:* RAVIN, Y. & LEACOCK , C. (eds.) *Polysemy. Theoretical and Computational Approach,.* Oxford,: Oxford University Press.

STOJANOVIC, N., MAEDCHE, A., STAAB, S., STUDER, R. & SURE, Y. 2001. A Framework for Developing SEmantic PortALs. *ACM K-CAP 2001.* Vancouver.

STUMPF, M., WIUF, C. & MAY, R. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *PNAS.* National Academy of Sciences USA.

SULLIVAN, D. 2001. *Document warehousing and text mining,* New York, Wiley Computer Publishing.

SUMMERFIELD, M. 2007. *Rapid GUI programming with Python and Qt : the definitive guide to PyQt programming,* Michigan. EU., Pearson Education, Inc.

SWARTIU, B., PATIL, R., KNIGHT, K. & RUSS, T. 1997. Toward distributed use of large-scale ontologies. *AAAI-97 Spring Symposium Series on Ontological Engineering.*

TENOPIR, C. & JACSÓ, P. 2007. *Quality of abstracs* [Online]. Santa Clara. Available: http://intranet.cdict.uclv.edu.c u [Accessed 26.marzo 2007].

TESNIÈRE, L. 1959. *Éléments de syntaxe structurale,* París, Klincksieck.

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

THEODORIDIS, S. & KOUTROUBAS, K. 1999. *Pattern Recognition,* Academic Press.

THOMPSON, S. 1991. *Methods of abstracting,* Cambridge, Cambridge University Press.

THOMPSON, S. & MANN, L. 1988. *Rethoric estructure in text*, McGraw-Hill.

TOMLIN, K. 2003. *Pragmatic and semantic in discurse*.

TOMLISON, D. M. 1986. LISA: anatomy of an abstracting service. *Indexer,* 15**,** 83-86.

TORRE, F. & KANADE, T. 2006. Discriminative cluster analysis. *In Proceeing of ICML '06: 23rd International Conference on Machine Learning.* Pensylvania: ACM Press.

TORRES-MORENO, J. M., VELÁZQUEZ-MORALES, P. & MEUNIER, J. G. 2002. Condensés de textes par des méthodes numériques. *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT).* St. Malo.

TURNER, P. D. n.d. Measuring Semantic Similarity by Latent Relational Analysis.

TUZHILIN, A. 2002. *Handbook of Data Mining and Knowledge Discover,* Oxford, Oxford University Press.

UNE 1990. *Documentación : preparación de resúmenes 50-103-90* Madrid, AENOR.

URÍAS, G. 2009. *Metodología de la Investigación,* Santa Clara, Universidad Central"Marta Abreu" de las Villas.

USCHOLD, M. & GRONINGER, M. 1996. Ontologies: Principles Methods and Applications. *Knowledge Engineering Review,* 2.

USCHOLD, M. & KING, M. 1995. Towards a Methodology for Building Ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing.*

USCHOLD, M. & KING, M. 1995. Towards a Methodology for Building Ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing*.

VAN DIJK, T. 1978. *La Noticia como discurso: comprensión, estructura y producción de la información,* Barcelona, Paidós.

VAN DIJK, T. 1980. *Estructura y funciones del discurso,* México, D.F., Siglo veintiuno.

VAN DIJK, T. 1984. *Texto y Contexto: Semántica y Pragmática del discurso,* Madrid, Cátedra.

VAN DIJK, T. 1995. De la gramática del texto al análisis crítico del discurso. *BELIAR,* 2.

VAN DIJK, T. 2004. *Discurso y desigualdad,* Tenerife, Universidad de La Laguna.

VAN DIJK, T. n.d. *La Iniciativa de Comunicación* [Online]. Available: http://www.comminit.com/la/lacth/sld-5183.html [Accessed 28.marzo 2009].

VAN DIJK, T. & KINTSCH, W. 1983. *Strategies of discourse comprehension,* Orlando, Fla., Academic Press.

VAN SLYPE, G. 1990. *El servicio de documentación frente a la explosión de la información,* Buenos Aires, Consejo Nacional de Investigaciones Científicas y técnicas.

VERMA, R. & CHEN, P. n.d. A Semantic Free-text Summarization System Using Ontology Knowledge.

VICKERY, B. 1997. Metatheory and Information Science. *Journal of Documentation,* 53**,** 457-476.

VILLASEÑOR, I. 1996. Las Fuentes de información. *In:* LÓPEZ YEPEZ, J. (ed.) *Manual de Información y Documentación.* Madrid: Edicines Pirámide.

VILLENA, J., GONZALEZ, J. & GONZALEZ, B. 2002. Dedalus: modelo de desambiguación.

VIRDHAGRISWARAN, S. 1994. Heterogeneous Information Systems Integration - an Agent Messaging Based Approach. *Third International Conference on Information and Knowledge Management (CIKM'94).* Gaithersburg, Maryland.: National Institute of Standards and Technology.

VIZCAYA, D. 1996. *Lenguajes documentarios,* Rosario, Argentina, Nuevo Paradigma.

WANG, W., YANG, J. & MUNTZ, R. 1997. STING: a statistical information grid approach tospatial data mining. *In Proceeding of 23rd International Conference on Very Large Data Bases.* Athens, Greece: Morgan Kaufmann.

WANG, Y. & YI-SHUN, W. n.d. Examining the dimensionality and measurement of user-perceived knowledge and information quality in the KMS context. *Journal of Information Science,* 35.

WASSERMAN, S. & FAUST, K. 1994. *Social network analysis: methods and applications,* Cambridge, Cambridge University Press.

WAYNE, C. n.d. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation

WEI, C., YANG, C. & LIN, C. 2008. A Latent Semantic Indexing-Based Approach to Multilingual Document Clustering. *Decision Support Systems,* 45**,** 606-620.

WHFAFLF, A. & ARMSTRON, C. 1997. Metadata, recall, and abstracts: can abstracts ever be reliable indicators of docuni iit value? . *Aslib*

WIEGAND, T. & MOLONEY, K. 2004. Rings, circles and null-models for point pattern analysis in ecology. *Oikos,* 104**,** 209-229.

WIELING, M. & NERBONNE, J. 2009. Bipartite Spectral Graph Partitioning to Co-Cluster Varieties and Sound Correspondences in Dialectology. *In:* CHOUDHURY, M., HASSAN, S., MUKHERJEE, A. & MURESAN, S. (eds.) *Text Graphs 4, Workshop at the 47th Meeting of the Association for Computational Linguistics.* Singapore.

WITTEN, I. & FRANK, E. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*.

XIE, X. L. & BENI, G. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Learning,* 13**,** 841-846.

XIONG, H., WU, J. & CHEN, J. 2006. K-means clustering versus validation measures: a data distribution perspective. *12th ACM SIGKDD international conference on Knowledge discovery and data mining.* Philadelphia: ACM Press.

XU, G. 2004. *Data mining,* Londres, Oxford University Press.

XU, X., YURUK, N., FENG, Z. & SCHWEIGER, T. 2007. TAJ. *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Jose, California.

YAGER, R. R. & PETRY, F. E. 2006. A Multicriteria Approach to Data Summarization Using Concept Ontologies. *IEEE TRANSACTIONS ON FUZZY SYSTEMS,* 14**,** 767-780.

YANG, Y. & PEDERSEN, J. 1997. A comparative study on feature selection in text categorization. *Journal of Artificial Intelligence Research,* 6**,** 1-34.

YANG, Z., ZHANG, D. & YE, C. 2006. Evaluation Metrics for Ontology Complexity and Evolution Analysis. *In:* IEEE (ed.) *International Conference on Advanced Language Processing and Web Information Technology.*

YAO, H., MARK, A., ORME, M. & ETZKORN, L. 2005. Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science,,* 1**,** 107-113.

VÉRONIS, J. 2002. Review of "Polysemy - Theoretical and computational approaches. *In:* RAVIN, Y. & LEACOCK., C. (eds.) *Computational Linguistics, .*

YAROWSKY, D. 1992. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of COLING-92.*

YAROWSKY, D. 1993. One Sense per Collocation. *DARPA Workshop on Human Language Technology.* Princeton, NJ, .

YAROWSKY, D. 1994. Decision Lists For Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. *Proceedings of the 32nd Meeting of the Association for Computational Linguistics.* Las Cruces, NM

YAROWSKY, D. 2000. Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* 34**,** 179-186

YEAL, W. & JEFFERIES, M. 1999. Computing a representation of the local environment. *Artificial Intelligence,* 107**,** 265-301.

YI, Y. & BEHESHT, J. 2009. A hidden Markov model-based text classification of medical documents *Journal of Information Science,* 35.

YING, Z. & KARYPIS, G. 2002. Evaluation of Hierarchical Clustering Algorithms for Documen: technical report. Minesotta: University of Minnesota, Department of Computer Science and Engineering.

YORK, U. D. N. 2002. *Proyecto Proteus* [Online]. Nueva York. Available: http://www.cs.nyu.edu/cs/projects/proteus/index.html [Accessed Julio.20 2007].

YOU, J., SANGHYUN, P. & INBUM, K. 2008. An efficient frequent melody indexing method to improve the performance of query-by-humming systems. *Journal of Information Science* 34.

YU, J., JAMES, A. T. & TAM, A. 2009. Requirements-oriented methodology for evaluating ontologies. *Information Systems Research,* 34**,** 766-791.

YUAN, S.-T. & SUN, J. 2004. Ontology-Based Structured Cosine Similarity in Speech Document Summarization. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)* IEEE.

ZAHN, C. 1971. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput,* 20**,** 68-86.

ZALDUA, A. 2006. El análisis del discurso en la organización y representación de la información-conocimiento: elementos teóricos. *ACIMED* 14.

ZAMORA-MARTÍNEZ, F. & CASTRO-BLEDA, M. J. 2008. Traducción automática basada en n-gramas conexionistas. *Procesamiento del Lenguaje Natural,* 45.

ZANG, Z., HUANG, Z., ZHANG, X. & 2010. Knowledge Summarization for Scalable Semantic Data Processing. *Journal of Computational Information Systems,* 6**,** 3893-3902

ZHANG, T., RAMAKRISHNAN, R. & LIVNY, M. 1996. BIRCH: An efficient data clustering method for very large databases. *In Proceeding of International Conference on Management of Data (SIGMOD).* Montreal, QB, Canada: ACM Press.

ZHANG, T., XU, D. & CHEN, J. 2008. Application-oriented purely semantic precision and recall for ontology mapping evaluation. . *Knowledge-Based Systems, ,* 21**,** 794-799.

ZHANG, X., CHENG, G. & QU, Y. 2007. Ontology Summarization Based on RDF Sentence Graph. *WWW 2007.* Banff, Canadá: ACM.

ZHAO, Y., ZHANG, C. & SHEN, Y. 2004. Clustering high-dimensional data with low-orderneighbors. *In Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence.* IEEE. Computer Society.


these things
types outside himself
called τὸ ἀρχέτυπον φῶς
times in Dionysius
hierarchia, II,
nominibus, I,
is not found
diversis quaestion

# Anexos

# ANEXOS

## ANEXO 1 GUIA DE OBSERVACION

La guía de observación que se presenta a continuación describe los procesos principales que lleva a cabo un resumidor para realizar extractos de textos científicos. Los parámetros que incluye esta Guía son los siguientes:

PROCESO DE LECTURA

Lectura del sumario

Lectura del resumen del autor.

Lectura de las palabras claves.

Lectura del texto del documento

- Introducción
- Metodología
- Discusión
- Resultados

ANALISIS DEL DOCUMENTO

- Extracción de oraciones.
- Delimitación de párrafos
- Lectura de imágenes.
- Cortes de la estructura retórica nivel estructural (micro y macro estructura.

REPRSENTACION

- Estrategias cognitivas
- Modelos memorísticos
- Elementos de percepción declarar cuantos
- Organización del conocimiento.
- Forma de Recuperación de la Información

FUENTES DE INFORMACION PARA CONFECCIONAR EL RESUMEN

- Revistas (Tipología )

## ANEXO 2 TEST DE EVALUACIÓN DE LA ONTOLOGÍA

Para la evaluación de Ontosatcol se propone una guía de Análisis de Contenido que pretende ayudar a la formulación organizada de los contenidos. Para desarrollar esta guía de observación los expertos no requieren de un entrenamiento previo para el desarrollo de la actividad. Ya han poseen conocimientos de la Ontología pues han participado de todas la etapas de diseño y composición de la misma.

Pídales a los expertos que revisen la ontología mediante el pluguin de visualización desarrollado para estos casos en el sistema Ontosatcol. Al final de la sesión se le pedirá a cada experto la planilla de anotación, donde registrará la información de cada categoría. Cada vez que el experto visualice cada categoría registrará en una planilla las frecuencias con que aparecen los errores o las clases que estima que están erróneas. Se otorgarán los valores. Siguientes a cada categoría: 5 (siempre), 4 (casi Siempre), 3 (regularmente), 2 (pocas veces) y 1 (nunca). Para hacer cuantitativos los datos será necesario calcular el % de las frecuencias de cada categoría.

**Población:** Los expertos que participan son los profesores del Departamento de Biología de la Universidad Central “Marta Abreu” de las Villas. Las unidades de análisis que incluye esta Guía son las siguientes:

Indicadores léxicos: Evalúan la cobertura terminológica del documentos, sus variables son: cobertura temática, exhaustividad de la anotación, redundancias semánticas, taxonomías ambiguas, capacidad de desambiguación y capacidad de traducción. Indicadores para la evaluación de la estructura sintáctica: Para ello se debe usar Protex, un pluguin que posee Ontosatcol, mediante el cual el experto logrará localizar y focalizar las siguientes variables.

**Indicadores de recuperación de información:** Se encarga de medir la capacidad de la ontología para recuperar la información. Las variables que miden estos indicadores son las siguientes: precisión y exhaustividad.

**Delimitación de las Categorías de Observación (presentación de las subcategoría de observación)**

**Indicadores léxicos:**

a) cobertura temática
b) exhaustividad de la anotación

c) redundancias semánticas
d) taxonomías ambiguas
e) capacidad de desambiguación
f) capacidad de traducción

**Indicadores de recuperación de información**:

**a)** precisión
**b)** exhaustividad.

**Indicadores para la evaluación de la estructura sintáctica**

a) Identificación de solapamiento en el desarrollo de clases.
b) Localización de conceptos vacíos.
c) Omisión de conocimiento disjunto.
d) Omisión de conocimiento por falta de exhaustividad.
e) Conceptos mal formulados.

Tabla de Recogida de Datos: Indicadores Léxicos

| Indicador / individuo | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |

Tabla de Recogida de Datos para: **Indicadores de recuperación de información**

| Indicador | a | b |
|---|---|---|

| individuo | | |
|---|---|---|
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |

***Tabla de Recogida de Datos para:***

Indicadores para la evaluación de la estructura sintáctica

| Indicador / individuo | a | b | c | d | e |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |
| 7 | | | | | |
| 8 | | | | | |
| 9 | | | | | |
| 10 | | | | | |

## ANEXO 3 GUIA OBSERVACION DE LAS CATEOGORÍA DE LOS CORPUS

Dos expertos en lingüística valorarán las cualidades del texto teniendo en cuenta los siguientes parámetros, los expertos registrarán los valores en la planilla que existe para estos efectos. Todas las coincidencias deben ir registradas en la planilla, a partir de dos análisis sobre el corpus inglés y el corpus español. Los resultados luego serán calculados con $Chi^2$. Los expertos conocen de antemano el valor de estas categorías y están previamente entrenados para hacer las valoraciones.

Las categorías que han de describirse son:

- Similaridad de Contenido
- Similaridad Estructural
- Similaridad Lenguaje
- Similaridad/imag.
- Overall Similarity

| Métricas | Cat.Iguales | | Cate.Diferentes | | Resultados | |
|---|---|---|---|---|---|---|
| Similiradidad de Contenido | A | B | A | B | A | B |
| Similaridad Estructural | | | | | | |
| Similaridad Lenguaje | | | | | | |
| Similaridad/imag. | | | | | | |
| Overall Similarity | | | | | | |

## ANEXO 4 DECLARACION DE ELEMENTOS QUE INDICAN RELEVANCIA E IRRELEVANCIA EN LOS TEXTOS DE LOS TEXTOS DEL CORPUS

Esta guía de análisis de contenido se realiza para que se declaren tanto aquellas unidades del texto que ofrecen relevancia semántica como aquellas que no ofrecen relevancia.

Las unidades de análisis para relevancia son las siguientes:

Topónimos: Términos geográficos

Fórmulas: Fórmulas Químicas

Compuestos Químicos: Mención a compuestos químicos

Palabras en Plural: Palabras declaradas en plural

Palabras en Singular: Palabras de consistencia semántica en singular

Frases: Frase de nivel sintáctico de elevado nivel connotacional

Combinaciones de Frases: Frase combinadas

Sustantivos: Sustantivos propios del dominio

Adjetivos: Adjetivos propios del dominio

Construcciones: Formaciones de adjetivos y Sustantivos

Registre mediante la siguiente tabla la presencia de los elementos que se declaran. Utilice la herramienta Word Smith Toolls para reconocer los términos que se corresponden con el dominio que se estudia. Debe también declarar las unidades nominales del discurso

| Elemento | % |
|---|---|
| Topónimos | |
| Lugares Geográficos | |
| Verbos | |
| Fórmulas | |
| Compuestos Químicos | |
| Palabras en Plural | |
| Palabras en Singular | |
| Frases | |
| Combinaciones de Frases | |
| Sustantivos | |
| Adjetivos | |
| Construcciones | |
| Unidades nominales | |

Para declarar los elementos que indican irrelevancia registre la información en esta tabla. Los datos de ambas tablas serán contabilizados y desarrollados en porciento.

| Elemento | Cantidad |
|---|---|
| Datos estadísticos o computacionales | |
| Análisis de Tablas y/o figuras | |
| Definiciones dudosas | |
| Datos Históricos | |
| Frases de escasa consistencia científica | |
| Trabajos previos o relacionados | |

## ANEXO 5 ANALISIS DE LA ESTRUCTURA DE LOS CORPUS

Declare en esta tabla las características que observa en el Corpus Inglés y en el español. Para ello deben estudiar las cualidades esenciales de estos textos. Declare los elementos deficientes en ambos corpus basándose en las siguientes variables: Presencia de resumen del autor, estructura científica, claridad en la redacción técnica.

**Variables:**

Presencia del Resumen del Autor

Estructura científica (IOMRC)

Claridad en la redacción técnica

Llene la tabla establecida para focalizar las características de los corpus. Dando un valor 0 para lo que considere totalmente desfavorable, 1 medianamente desfavorable y 2 Favorable.

| Variables | 0 | 1 | 2 |
|---|---|---|---|
| Presencia del Resumen del Autor | | | |
| Estructura científica (IOMRC) | | | |
| Claridad en la redacción técnica | | | |

## ANEXO 6 GUIA DE ANÁLISIS DE CONTENIDO PARA EL ESTUDIO DEL DISCURSO EN INGENIERÍA DE PUERTOS Y COSTAS

Esta guía de observación se realiza con el objeto de servir de instrumento de análisis de los textos. Los textos que se analizarán oscilarán entre 50 y 10 de acuerdo al nivel textual que se analice. Esta guía es necesaria para normalizar y registrar el estudio del corpus español del proyecto PUERTOTERM. Este procedimiento lleva del entrenamiento de varia semanas y se realiza con el concurso de registradores expertos en lingüística que revisan la calidad del análisis.

Las **Variable**s a analizar son las siguientes:

**Análisis Semántico:** Registro del significado global del texto mediante ejes semánticos.

**Subcategorías**

**Procesos:** Declaran actividades inherentes a la naturales o al hombre.

**Agentes:** Hacen influencia sobre alguna entidad.

**Medición:** Formas de medición de algunas actividades.

**Aplicación:** Aplicación, donde se aplica el proceso

**Efectos:** Lo que ocurre al aplicar algo o lo que sucede cuando pasa algo, que puede ser fenómeno, acción del hombre, etc.

**Técnicas:** Cuáles son las técnicas de análisis

**Representación:** Forma en que se representa algún fenómeno.

**Muestras:** Para ello se deben estudiar 50 textos del corpus de PUERTOTERM.

**Procedimiento:**

- Lectura del documento o texto
- Divida el texto en unidades retóricas o segmento (Objetivos, Metodologías, Discusión y Resultados).
- Seleccione las temáticas del texto utilizando las palabras clave que identifica cada grupo de oraciones, párrafo o sección.
- Localice en él elementos de unión o ejes tipológicos como parte de, entre, etc.
- Registre la Información en una tabla donde aparezca la frecuencia con que aparezca cada repetición de los ejes semánticos.

Este procedimiento se realizará con un intervalo de 40 minutos para cada texto y se entregarán los textos graficados y registrados a los registradores.

**Herramienta:** La herramienta utilizada será Cmaptooll, indicada en un principio para mapa conceptual pero muy útil para desarrollar y describir contenidos.

**Registro de los Datos**: Registre en una Tabla diseñada a efecto la frecuencia en que los ejes aparecen en los textos, si apareciera otro declárelo en la planilla al dorso de la misma en la sección adiciones.

**Tabla para la Recogida de Datos**

| **Subcategorías** | **Text 1** | **Text 2** | **Text 3** | **Text 4** | **Text 5** |
|---|---|---|---|---|---|
| Procesos | | | | | |
| Agentes | | | | | |
| Medición | | | | | |
| Aplicación | | | | | |
| Efectos | | | | | |
| Técnicas | | | | | |
| Representación | | | | | |
| **Subcategorías** | **Text 6** | **Text 7** | **Text 8** | **Text 9** | **Text 10** |
| Procesos | | | | | |
| Agentes | | | | | |
| Medición | | | | | |
| Aplicación | | | | | |
| Efectos | | | | | |
| Técnicas | | | | | |
| Representación | | | | | |
| **Subcategorías** | **Text 11** | **Tex 12** | **Text 13** | **Tex 14** | **Tex 15** |
| Procesos | | | | | |
| Agentes | | | | | |
| Medición | | | | | |
| Aplicación | | | | | |
| Efectos | | | | | |
| Técnicas | | | | | |
| Representación | | | | | |
| **Subcategorías** | **Text 16** | **Tex 17** | **Tex 18** | **Text 19** | **Text20** |
| Procesos | | | | | |
| Agentes | | | | | |
| Medición | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Aplicación | | | | | |
| Efectos | | | | | |
| Técnicas | | | | | |
| Representación | | | | | |

**Macro estructura:** Registro de las Unidades Macroestructurales del texto, o sea los elementos en que el texto aparecen de forma global.

**Oraciones de Síntesis**: Aparecen y su posición en apartados del texto
**Macro Reglas:** Elisión de oraciones de bajo nivel semántico y la generalización para la aglutinación de ideas y conceptos

**Muestras:** Para ello se deben estudiar 50 textos del corpus de PUERTOTERM.

**Procedimiento:**

- Lectura del documento o texto
- Divida el texto en unidades retóricas o segmentos macroestructurales (Objetivos, Metodologías, Discusión y Resultados).

**Registro de los Datos**: Registre en una Tabla diseñada al efecto la frecuencia los elementos Macroestructurales del texto.

| **Macroestructura** | Text 1 | Text 2 | Text 3 | Text 4 |
|---|---|---|---|---|
| Introducción | | | | |
| Metodología | | | | |
| Objetivos | | | | |
| Resultados | | | | |
| Conclusiones | | | | |

**Registre en el qué textos aparecen estos apartados**

| **Oraciones de Síntesis** | Posición final | Posición intermedia | Posición inicial | Cierre conclusivo |
|---|---|---|---|---|
| Text 1 | | | | |
| Text 2 | | | | |
| Text 3 | | | | |
| Text 4 | | | | |
| Text 5 | | | | |
| Text 6 | | | | |
| Text 7 | | | | |

| | | | | |
|---|---|---|---|---|
| Text 8 | | | | |
| Text 9 | | | | |
| Text 10 | | | | |

**Registre cuantas oraciones están declaradas en cada posición**

| **Oraciones de Síntesis** | Bajo Nivel Semántico | Bajo Nivel de Generalización |
|---|---|---|
| Text 1 | 1 | |
| Text 2 | 2 | |
| Text 3 | 3 | 2 |
| Text 4 | 4 | 4 |
| Text 5 | 5 | 6 |
| Text 6 | 3 | |
| Text 7 | 6 | |
| Text 8 | 1 | 1 |
| Text 9 | | 2 |
| Text 10 | | 2 |
| Text 11 | | |
| Text 12 | | |
| Text 13 | | |
| Text 14 | | |
| Text 15 | | |
| Text 16 | | |
| Text 17 | | |
| Text 18 | | |
| Text 19 | | |
| Text 20 | | |
| Text 21 | | |
| Text 22 | | |
| Text 23 | | |
| Text 24 | | |
| Text 25 | | |
| Text 26 | | |
| Text 27 | | |
| Text 28 | | |
| Text 29 | | |

| | | |
|---|---|---|
| Text 30 | | |
| Text 31 | | |
| Text 32 | | |
| Text 33 | | |
| Text 34 | | |
| Text 25 | | |
| Total | 25 | 18 |

**Registre cuáles oraciones poseen bajo nivel semántico en el texto y cuáles poseen bajo nivel de generalización.**

**Variable Fuerza Semántica**

Para localizar la fuerza semántica de las oraciones dentro de los apartados es necesario poner en qué posición del texto se encuentran para declarar regularidades. Para ello es imprescindible tomar cada apartado para decir en qué lugar del mismo se colocan las oraciones de mayor nivel semántico.

Se debe registrar la posición de cada oración en cada apartado en la tabla que se muestra. La cantidad de resúmenes a analizar son 50.

| Elemento estructural | 3 última y 3 primeras | 3 primeras y 4 últimas oraciones | 3 últimas oraciones | Junto a metodología | 5 primeras o 3 últimas oraciones |
|---|---|---|---|---|---|
| Introducción | 0 | 89 % | 4 % | 6% | 0 |
| Metodología | 82% | 10 % | 8 % | 0 | 0 |
| Resultados | 6% | 5,0 % | 78,9 % | 10, % | 0 |
| Conclusiones | 0 | 0 | 0 | 10 | 90% |
| Objetivos | 0 | 0 | 16%6 | 67 % | 17 % |

**ANEXO 7 ESTUDIO DE NECESIDADES**

Con el objetivo de estudiar las necesidades de la Comunidad Ciencias Biológicas se desarrolla este cuestionario para el estudio de necesidades, en el mismo el usuario debe declarar las siguientes variables.

Datos Generales:

- Nombre y Apellidos
- Sexo
- Facultad

Temáticas en las que Investiga

Fuentes de Información

- Revistas
- Bases de Datos
- Autores
- Libros
- Congresos
- Conferencias

Funcionamiento del Sistema

- Hora en que necesita el Servicio
- Forma en que necesita la información: Textual, Gráfica

Idioma en lee los textos

**ANEXO 8 CUESTIONARIO DE ADECUACION DE LOS REQUERIMIENTOS**

Estimado usuario para verificar los resultados de evaluación de la Ontología se necesita adecuar los requerimientos, respondiendo este test de usabilidad localizando en la ontología los elementos mediante preguntas. La variable a analizar se denominará Recuperación de la Información. El resultado de la aplicación de las preguntas se llevará a escala de puntuación para su calificación con una escala de valores de “si”, “a veces” y “nunca”. Cada evaluador hará la búsqueda 3 veces y anotará las veces que recupera la información. Los datos se llevaran a nivel porcentual y se registrarán de forma individual por cada evaluador, finalmente se tabulará y se presentarán en datos homogéneos.

Preguntas:

| No. | Pregunta | Si | A veces | No |
|---|---|---|---|---|
| 1 | ¿Cuáles son los efectos que produce la marea en la Costa de Cantabria? | | | |
| 2 | ¿Cuáles son las centrales azucareras más contaminantes de Cuba? | | | |
| 3 | ¿Con qué sustancia química se hace la espectrografía de gases? | | | |
| 4 | ¿Cuáles son las zonas geográficas de menor desarrollo portuario? | | | |
| 5 | ¿Cuál es la traducción del término “rompiente de derrame” en alemán? ¿Cuál es su contexto de aplicación? | | | |

| | | | | |
|---|---|---|---|---|
| 6 | ¿Cuántas imágenes de derrame de petróleo existen en la ontología? | | | |

**ANEXO 9 TEST DE USABILIDAD**

Usuario con el objeto de realizar pruebas de usabilidad en el sistema le pedimos realice 8 tareas de trabajo con el sistema. Para cada tarea hay un margen de 20 minutos. Debe marcarse con 0 las tareas no cumplidas, con un 1 aquellas conseguidas con dificultades y con un 2 las que ha realizaron fácilmente. La variable de estudio es recuperación de la información. A continuación se declaran las tareas que han de acometerse con el sistema.

1. Usted trabaja en la Facultad de Biología de la UCLV y necesita buscar el término EMBALSE en INGLES. Localícelo en el sistema.
2. Busque las posibles consecuencias de las mareas altas.
3. Busque cuáles son las tipos de bahía que existen en España.
4. Localice un resumen del autor Polioptro Machado Randín.
5. Localice información sobre la junta de Andalucía. (En este caso no buscamos palabras equivalentes sino información sobre un tema.)
6. Localice una imagen de la bahía de Cádiz.
7. Envíe al administrador su opinión sobre el servicio de búsqueda de PUETRTOTEX.
8. Entre al sistema y compruebe que lenguas tiene disponibles para los usuarios.

| No. | Acción | 0 | 1 | 2 |
|---|---|---|---|---|
| 1 | Usted trabaja en la Facultad de Biología de la UCLV y necesita buscar el término EMBALSE en INGLES. Localícelo en el sistema | | | |
| 2 | Busque las posibles consecuencias de las mareas altas | | | |
| 3 | Busque cuáles son las tipos de bahía que existen en España | | | |
| 4 | Localice un resumen del autor Polioptro Machado Randín | | | |
| 5 | Localice información sobre la junta de Andalucía. (En este caso no buscamos palabras equivalentes sino información sobre un tema.) | | | |

| | | |
|---|---|---|
| 6 | Localice una imagen de la bahía de Cádiz | |
| 7 | Envíe al administrador su opinión sobre el servicio de búsqueda de PUETRTOTEX | |
| 8 | Entre al sistema y compruebe que lenguas tiene disponibles para los usuarios. | |

## ANEXO 10 CUESTIONARIO DE EVALUACIÓN DE SISTEMA

Con el objetivo de Evaluar el sistema PUERTOTEX, le pedimos que analice las siguientes preguntas y que responda acorde a su experiencia con el sistema. Marque con una cruz el valor que le concedes a la variable.

| Interrogantes | Sí | A veces | *No* |
|---|---|---|---|
| ¿El sistema de navegación permite al usuario la Recuperación efectiva de la Información? | | | |
| ¿Están declaradas en el sistema todas sus funcionalidades? | | | |
| ¿Puede controlar el sistema con facilidad? | | | |
| ¿El sistema expresa claramente la lengua y el contenido? | | | |
| ¿Ofrece el sistema al usuario Ayuda en línea? | | | |
| ¿Existe diferencia entre las interfaces de búsqueda y recuperación en el sistema? | | | |
| ¿Es accesible el sistema? | | | |
| ¿Es Coherente la presentación de las tablas? | | | |
| ¿El sistema está hecho para la Prevención errores? | | | |
| ¿Posee el sistema claridad arquitectónica? | | | |

## ANEXO11 EVALUACION POR EXPERTOS DEL MODELO TEXMINER

Sr(a): Me llamo Amed Abel Leiva Mederos, estudiante de Doctorado de la Universidad de Granada, mediante este texto le invito a evaluar el modelo de construcción automática de resúmenes que propongo en mi tesis, por ser usted uno de los expertos con elevada calificación en el tema. La lectura del modelo solo exigirá de unos 15 minutos y cumplimentar la encuesta también. Le agradecería que brindara sus observaciones y valoraciones sobre la propuesta.
**Nombre**: TEXMINER. Modelo para la extracción y desambiguación de textos científicos.
El modelo que se presenta tiene como ***objetivo general***: Construir resúmenes automáticos a partir de textos científicos.
Dentro de las ***características generales*** del modelo son:

7. Estudio de Necesidades.
8. Análisis Manual del Corpus Textual.
9. Creación de la ontología.
10. Extracción del Texto mediante Agentes Cognitivos.
11. Modelación de sistema de búsqueda y recuperación de información.
12. Representación de la Información

Teniendo en cuenta estas particularidades, se establecieron los elementos necesarios para la construcción del modelo, considerando como dispositivos imprescindibles para lograr su efectividad a los procesos que aquí se describen. La teoría aquí planteada emana de de los diversos instrumentos y métodos desarrollados en la investigación:

- Examinar los referentes teóricos metodológicos que han afectado al resumen automático y su desarrollo, haciendo énfasis en las preposiciones que han aparecido en Ciencias de la Computación, Ciencias de la Información, la Lingüística, la Semiótica y la Cibersemiótica.
- Estudio de las estrategias cognitivas de 12 resumidores para establecer regularidades en el proceso de resumir.
- Estudiar el dominio textual a través del análisis de discurso para establecer ejes semánticos.

- Construcción de reglas textuales a partir de las regularidades establecidas en los textos e implementar dichas reglas en agentes o robots que simulen estas regularidades.
- Evaluar la calidad de los textos a través de herramientas de análisis estadístico y con el criterio de los especialistas en lingüística.
- Utilizar herramientas de visualización que permitan la observación del contexto de cada concepto.
- Permitir la evaluación de los textos resultantes y el acceso a los textos originales, dotando al sistema de una real y lógica representación textual.
- Construir un sistema de extracción de textos que valide en la práctica el modelo.
- Evaluar el sistema teniendo en cuenta criterios de usabilidad.

A partir de las propuestas teóricas y procedimentales desarrolladas se determinaron los siguientes componentes para el modelo:

. ***Profesionales encargados del proceso***.

Este rol es importante en el modelo, ya que un sistema que trabaje con texto y con técnicas de inteligencia artificial necesita de un personal heterogéneo, por esta razón para ello se escoge:

1. **Bibliotecario:** Con una preparación que rebase los principios de la catalogación y la representación ordinaria, asumiendo formas de descripción estructurada, entiéndase por esto sistemas de marcado xml, estructuración semántica de contenidos, etc.
2. **Ingeniero Informático:** Uno de los errores en los procesos de desarrollo de aplicaciones para el tratamiento de información ha sido dejar en el terreno de los informáticos múltiples aplicaciones. En este caso el papel del informático es programar las operaciones cognitivas que detecten los bibliotecarios a través de las actualizaciones del modelo.
3. **Usuarios:** Juegan un papel fundamental en el desarrollo del sistema y del modelo, sus estrategias cognitivas son la base de los procesos que se automatizan y hacia ellos va dirigido el producto final: el resumen automático.

Todos los procesos de representación visual son desarrollados a partir de las necesidades de la comunidad de usuarios.

Estos roles constituyen una relación indispensable que no distingue de niveles en el modelo, demostrando la indisoluble unión entre los que revisan los contenidos, los que los distribuyen y quienes los consumen o usan.

***II. Recursos***

Por esta categoría se entiende todos los medios, los métodos y procedimientos que tributen al modelo. Esto propició reconocer elementos que destacan la relación biblioteca, organización y usuarios de la información.

Para lograr esta relación se necesitan, los planes de desarrollo de las organizaciones en los que se reflejen diversos datos como: , sus publicaciones, sus profesionales, así como el efectivo registro de necesidades de usuarios, tanto individuales como colectivos, además también se analiza el desarrollo de competencias informacionales tanto para usuarios como para todo el personal con roles en el modelo. Otros recursos que demanda el modelo son los tecnológicos, entre los que se encuentran máquinas y redes de computadoras.

***III. Interacción entre los componentes***

Todos los componente están relacionados, tanto a nivel procedimental (recogida de información) como técnico (unión correcta de las aplicaciones), cognitivo (estructuración del conocimiento). TEXMINER es el resultado de la interacción de los últimos paradigmas de la Representación de la información y el Conocimiento, específicamente el sociocognitivismo, que obliga a determinar las especificidades de la realidad representacional de cada usuario para utilizarla en los modelos de tratamiento de información, alejando los presupuestos de la organización y el resumen del paradigma físico, sustentado en la lingüística clásica para entrar en el terreno de la representación conceptual, más ligada a las vivencias de las comunidades de práctica. La interdependencia entre los componentes puede ser valorada a partir de los siguientes principios:

- La calidad de los resúmenes que ofrece a los usuarios.
- La capacidad para visualizar conceptos a partir de estrategias cognitivas.
- Descripción y Catalogación de Recursos a partir de modelos de metadatos más flexibles

- Búsqueda de información a partir de la semántica que declaren los conceptos
- La generación de redes sociales a partir de dominios específicos, lo que permitirá la comunicación interprofesional.

**Pasos del Modelo**


El gráfico declara las etapas del modelo, las mismas están enumeradas dando a entender la estructuración holística de todas las etapas, de forma que se genera un ciclo donde después de la evaluación del resumen se declaren nuevas necesidades de información que estén sustentadas en nuevas estrategias cognitivas. Hay por ende un modelo holístico para el desarrollo del resumen declarando aquellos elementos imprescindibles que desde la teoría de la Ciencia de la Información se ha exigido para estos materiales.

Cuestionario

De los principios del Modelo TEXMINER exprese su evaluación a través de los siguientes valores

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |

De los pasos declarados en la concepción teórica del modelo TEXMINER exprese su valoración a través de los siguientes indicadores.

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |

Realice cualquier valoración que considere sobre el modelo propuesto

Sobre la integración de los componentes en el modelo responda:

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |

Sobre los profesionales integrados al modelo responda:

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |

## ANEXO 12 CONTABILIZACIÓN DE LAS REGULARIDADES EN QUE APARECEN LAS REGLAS SINTÁCTICO COMUNICATIVAS.

Esta guía de análisis sirve para recoger las frecuencias en que aparecen las relaciones discursivas, para ello hay que marcar las frecuencias en que aparece cada una de las relaciones desde el punto de vista de relación, núcleo y satélite.

Para contabilizar estos elementos se necesita declarar la aparición de cada uno de los casos, que ya han sido declarados y validados por (D´Cunha, 2006)
A continuación se muestran las tablas donde se debe contabilizar las regularidades sintácticas comunicativas.

| Regularidad | Elemento Cohesivo | Frecuencia | % |
|---|---|---|---|
| **Regularidades en las que se elimina un satélite discursivo** | Concesión.<br>Reformulación.<br>Resultado.<br>Justificación.<br>Circunstancia.<br>Propósito. | 25 | 50% |
| **Regularidades en las que se elimina un núcleo discursivo** | núcleo de Interpretación<br>núcleo de Evidencia | 20 | 45% |
| **Regularidades en las que no se separa el satélite de su núcleo** | Satélite de Condición<br>Satélite de Resumen | 20 | 45% |
| **Regularidades en las que no se separan dos núcleos** | Lista<br>Contraste<br>Unión | 25 | 50% |

| | | | |
|---|---|---|---|
| **Regularidades en las que se elimina un satélite discursivo relacionado con un elemento sintáctico** | Appenditivo | 50 | 100% |
| **Regularidades en las que se elimina un satélite discursivo relacionado con un elemento comunicativo** | (Tema)<br>Background, Concesión, Reformulación, Resultado, Justificación Circunstancia | 50 | 100% |
| **Regularidades en las que no se elimina un satélite discursivo relacionado con un elemento comunicativo** | Elaboración(el contenido de un satélite discursivo de Elaboración se refiere al Rema de su núcleo) | 50 | 100% |
| **Reglas que eliminan elementos sintáctico-comunicativos** | Apenditivos se prefieren a Coordinativos | 50 | 100% |
| **No se suprimen aquellos satélites** | Satélites de Elaboración correspondientes a elementos | 50 | 100% |

| | | | |
|---|---|---|---|
| **discursivo** | atributivos | | |
| Regularidades de desambiguación léxica | Conjugar todos los verbos de primera persona a voz pasiva<br>Todas la referencias anafóricas del texto deben estar en línea de cohesión (esto se aplica en la base de conocimientos). | 50 | 100% |

## ANEXO 13 GUIA DE ENTREVISTA PARA LOS PSICÓLOGOS Y PEDAGOGOS

Esta guía de entrevista está destinada a los profesionales encargados de observar a los bibliotecarios en la redacción y extracción de textos especializados en ingeniería, la moderadora de la entrevista es la Dra. Graciela Urías Arboláez, vice-decana docente de la Facultad de Ciencias Sociales de la Universidad Central"Marta Abreu" de las Villas.

A continuación se listan las preguntas que la Dra hará a los profesionales:

1. ¿Cuáles son las habilidades cognitivas que lleva implícito el acto de resumir?
2. ¿Cuáles serían las variables que utilizaríamos en este estudio?
3. ¿Cómo creen que se debe realizar la evaluación de estas cualidades en la tesis?
4. ¿Qué elementos serían necesarios afianzar para no perder información?
5. ¿Qué importancia le atribuyen a la automatización de los reglas cognitivas en las comunidades de práctica?

## ANEXO 14 RESULTADO DE LA APLICACIÓN DE LA ENTREVISTA GRUPAL

***Presentación***

*Buenas tares, yo me llamo Graciela Urías Arboláez y voy a servir de moderadora en este ejercicio académico para determinar los indicadores más efectivos para la realización de la guía de observación que ha de ser la técnica inicial de recogida de información en el modelo TEXMINER.*

***Exposición de los Fundamentos***

***Graciela Urías Arboláez***

*Sobre este tema no hay mucha experiencia. La cibernética ha hecho estudios que aprovechan más la matemática, los lingüistas por su parte han declinado de estas actividades ya que prefieren asentar sus teorías en los valores de los Corpus. Hay una investigación que me ha llamado mucho la atención, es la desarrollada por la investigadora Brigitte Endres Niggemeyer en Alemania con el estudio de las estrategias cognitivas de los resumidores de artículos médicos. Ella en su estudio anotaba aquellas estrategias que más se ajustaban a la escritura y al resumen, incluso fue más allá, ella observó a los resumidores en su casa y en otras actividades. Yo se que todos se han leído el método y que tienen muchas cosas que decir. Yo los invito a expresar todos criterios sobre este tema en forma organizada.*

**Criterios de la Pregunta 1**

**Dra. Inés María Mederos Morell**

*Ya se ha estudiado este fenómeno de los procesos cognitivos. Desde el punto de vista pedagógico cuando hablamos de procesos cognitivos hay que hablar también de habilidades para la escritura, así que me parece que hay que empezar por determinar cuáles son las habilidades de trascripción y de redacción de los evaluados. El estudio de los alemanes fue muy efectivo, pero no declaran como lograron segmentar los procesos. Desde mi experiencia docente estos procesos empiezan por la lectura, seguidamente asumen la selección de oraciones, ya sea mediante el subrayado o la delimitación, luego se intenta normalizar y dar coherencia al texto y luego se hace una lectura secundaria para comprobar la calidad del texto.*

**Criterios de la Pregunta 2**

**Lic. Alibet Viera Muñoz**

*Concuerdo con la doctora que sería muy bueno determinar los procesos a partir de habilidades cognitivas, yo propongo que los procesos que se observen sean los siguientes:*

- *Lectura*
- *Subrayado*
- *Memorización*
- *Comprobación o Constatación*
- *Exploración del original*
- *Evaluación*
- *Memorización*
- *Tipos de Lectura*

*Esto puede enriquecerse con el criterio de todos los presentes.*

**Criterios de la Pregunta 3**

**MSc. Aymé Duquesne Morell**

*Esas cuestiones hay que registrarlas en una planilla, pero hay que darle tiempo a los compañeros, hay que decir que cada día cuando terminen las sesiones dejen en el vestíbulo los resúmenes y que declaren en vos alta las acciones cognitivas que van realizando, y hay que tener también un observador por cada sujeto, de lo contrario compañeros perderemos mucha información.*

**Criterios de la Pregunta 4**

**MSc. María Aleida Hernández**

*Creo que hay delimitar bien que se va a recoger, este ejercicio es muy complejo, personalmente hace mucho tiempo que no tenía un test tan complejo. Propongo que se delimiten las estrategias de las habilidades, es mejor que hablemos de estrategias solo y que nos olvidemos por el momento de las habilidades ya que esto es otro tipo de investigación.*

**Criterios de la Pregunta 5**

**Lic. Ariel González**

*Creo que es importante este estudio, su complejidad es innegable pero su utilidad en el desarrollo de los estudios cognitivos facilitará la construcción de la*

*estrategia de análisis de los contenidos que más tarde será estrategia de resumen.*

## ANEXO 15 GUIA DE ANÁLISIS DE LA CALIDAD DE LOS CORPUS DESPUES DEL RESUMEN

Expertos, en esta guía estamos declarando aquellas variables que tendrán que analizar para la valoración del contenido de los resúmenes obtenidos automáticamente mediante el modelo TEXMINER. Para el desarrollo de la guía solo basta con analizar 10 textos que fueron resumidos con el resumidor PURTOTEX y valorar en ellos la coherencia, la cohesión y el balance textual. Como ustedes poseen elevado nivel dentro del tema solo nos limitaremos en esta guía a describir las variables y la tabla que debe recoger el análisis de cada texto.

**Variables:**

Coherencia: Establecimiento lógico de las unidades textuales

Cohesión: Grado de unidad entre los grupos oracionales

Balance Textual: Separación correcta de grupor oracionales

| Variable | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cohesión | | | | | | | | | | |
| Coherencia | | | | | | | | | | |
| Balance | | | | | | | | | | |
| Textual | | | | | | | | | | |

Se les recomienda a los expertos que asienten las cualidades de cada texto de forma individual.

**ANEXO 16 GUIA DE ANÁLISIS DE LA CALIDAD DE LOS CORPUS DESPUES DEL RESUMEN (RESULTADOS)**

Expertos, en esta guía estamos declarando aquellas variables que tendrán que analizar para la valoración del contenido de los resúmenes obtenidos automáticamente mediante el modelo TEXMINER. Para el desarrollo de la guía solo basta con analizar 10 textos que fueron resumidos con el resumidor PURTOTEX y valorar en ellos la coherencia, la cohesión y el balance textual. Como ustedes poseen elevado nivel dentro del tema solo nos limitaremos en esta guía a describir las variables y la tabla que debe recoger el análisis de cada texto.

**Variables:**

*Coherencia:* Establecimiento lógico de las unidades textuales

*Cohesión:* Grado de unidad entre los grupos oracionales

*Balance Textual:* Separación correcta de Unidades Oracionales

| Variable | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cohesión | X | X | X | X | x | x | X | x | X | x |
| Coherencia | x | | X | x | X | X | X | | | X |
| Balance Textual | | X | X | | x | X | | X | x | X |

Se les recomienda a los expertos que asienten las cualidades de cada texto de forma individual marcando con una cruz el nivel de calidad

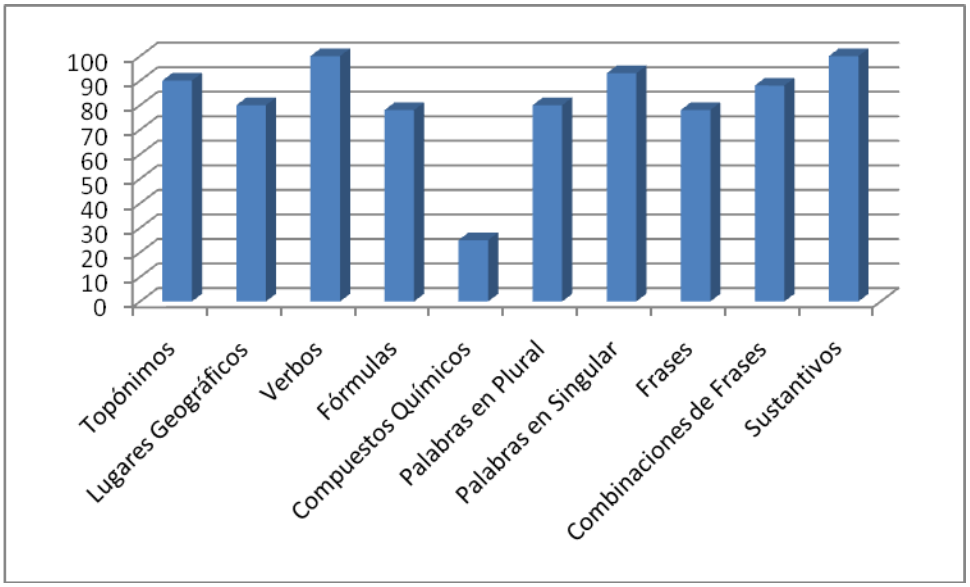
## ANEXO 17 RESULTADOS DE LA DECLARACION DE ELEMENTOS QUE INDICAN RELEVANCIA E IRRELEVANCIA EN LOS TEXTOS DE LOS TEXTOS DEL CORPUS

Esta guía de análisis de contenido se realiza para que se declaren tanto aquellas unidades del texto que ofrecen relevancia semántica como aquellas que no ofrecen relevancia.

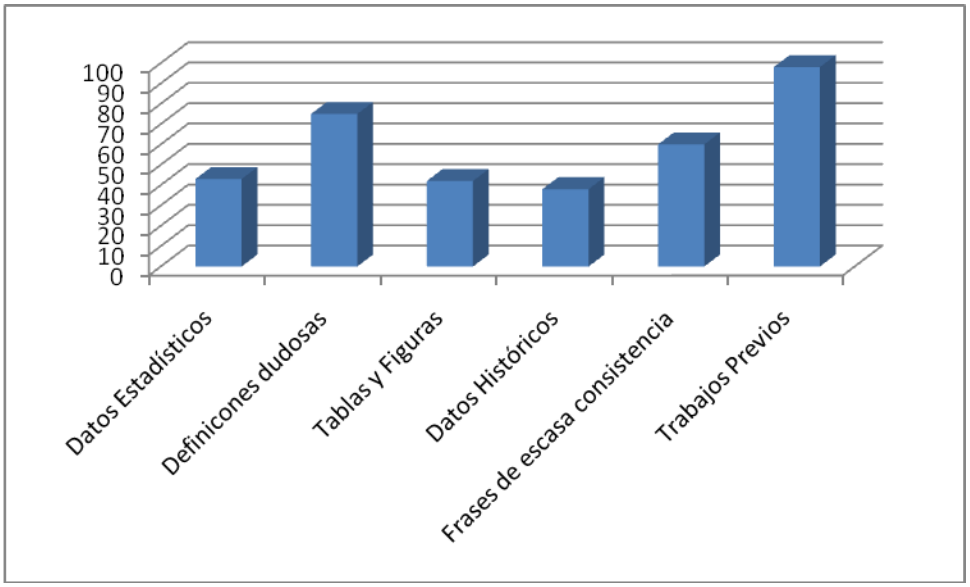
Las unidades de análisis para relevancia son las siguientes:

**Topónimos:** Términos geográficos
**Fórmulas**: Fórmulas Químicas
**Compuestos Químicos:** Mención a compuestos químicos
**Palabras en Plural:** Palabras declaradas en plural
**Palabras en Singular:** Palabras de consistencia semántica en singular
**Frases:** Frase de nivel sintáctico de elevado nivel connotacional
**Combinaciones de Frases:** Frase combinadas
**Sustantivos:** Sustantivos propios del dominio
**Adjetivos:** Adjetivos propios del dominio
**Construcciones:** Formaciones de adjetivos y Sustantantivos

Registre mediante la siguiente tabla la presencia de los elementos que se declaran. Utilice la herramienta Word Smith Toolls para reconocer los términos que se corresponden con el dominio que se estudia. Debe también declarar las unidades nominales del discurso




Para declarar los elementos que indican irrelevancia registre la información en esta tabla. Los datos de ambas tablas serán contabilizados y desarrollados en porciento.



**ANEXO 18 SENTIDO DE LA VOZ ELOCUTIVA EN LOS RESÚMENES**

Este Anexo tiene el objeto de determinar el modo en que se describen los textos. Para ello habrá que analizar los 50 textos para determinar en cuál de las formas de escritura está escrito.

A continuación se declaran las variables de análisis:

Vos pasiva
Primera Persona
Tercera Persona

| Texto | Vos Pasiva | Primera Persona | Tercera Persona |
|---|---|---|---|
| 1 | X | | |
| 2 | X | | |
| 3 | X | | |
| 4 | X | | |
| 5 | X | | |
| 6 | X | | |
| 7 | X | | |
| 8 | X | | |
| 9 | X | | |
| 10 | X | | |
| 11 | X | | |
| 12 | X | | |
| 13 | X | | |
| 14 | X | | |
| 15 | X | | |
| 16 | X | | |
| 17 | X | | |
| 18 | X | | |
| 19 | X | | |
| 20 | X | | |
| 21 | X | | |
| 22 | X | | |

| | | | |
|---|---|---|---|
| 23 | X | | |
| 24 | X | | |
| 25 | X | | |
| 26 | | X | |
| 27 | | X | |
| 28 | | X | |
| 29 | | X | |
| 30 | | X | |
| 31 | | X | |
| 32 | | X | |
| 33 | | X | |
| 34 | | X | |
| 35 | | X | |
| 36 | X | x | |
| 37 | X | | |
| 38 | X | | |
| 39 | X | | |
| 40 | X | | |
| 41 | X | | |
| 42 | X | | |
| 43 | X | | |
| 44 | X | | |
| 45 | X | | |
| 46 | X | | |
| 47 | X | | |
| 48 | X | | |
| 49 | X | | |
| 50 | X | | |

**ANEXO 19 GUIA DE ANÁLISIS DE UNIDADES LÉXICAS DE LOS ARTÍCULOS DE INGENIERÍA DE PUERTOS Y COSTAS**.

De los 50 textos, extraiga las siguientes variables: 1unidades léxicas nominales, 2 unidades léxicas verbales, 3 unidades léxicas del título principal. Asiente los resultados en el modelo que aparece al a continuación:

| Var. | t1 | t2 | t3 | t4 | t6 | t7 | t8 | t9 |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |

Debe poner el nombre de la unidad nominal que encuentre en la casilla del texto correspondiente. Para esto se debe usar la herramienta Wordsmith toolls.

**ANEXO 20 GUIA DE ANALISIS DE LOS ELEMENTOS COHESIVVOS DEL TEXTO.**

Para analizar los elementos cohesivos del texto debe leer todos los textos y en cada oración marcar las estructuras cohesivas de los textos y las posiciones en que se encuentran todos los núcleos y los satélites de 50 textos.

| |
|---|
| **Contraste (M)** |
| **Unión (M)** |
| **Lista (M)** |
| **Secuencia (M)** |
| **Backgroug (N-S)** |
| **Concesión (N-S)** |
| **Condición (N-S)** |
| **Elaboración (N-S)** |
| **Justificación (N-S) y (M)** |
| **Propósito (N-S)** |
| **Reformulación (N-S)** |
| **Resultado (N-S)** |
| **Resumen (N-S)** |
| **Evidencia (N-S)** |
| **Interpretación (N-S)** |

**ANEXO 21 CONTEO DE LOS TIPOS DE VOCABLOS Y DE LA TIPOLOGIA DISCURSIVA EN LOS CORPUS DE LA INVESTIGACION**

Utilizando Wordsmith y Protex y las ontologías Wordned y Euroword ned declare los elementos que tienen categoría que se enuncian en la tabla.

**Frecuencia de Términos**

| **Corpus** | verbos | adverbios | homónimos | merónimos | hipónimos | **Total** |
|---|---|---|---|---|---|---|
| C.Español | | | | | | |
| C.Inglés | | | | | | |

Analizando los Artículos de ambos corpus determine los que se corresponden con las variables que se ponen en la tabla.

**Tipología Discursiva**

| **Corpus** | Artículos Científicos | Artículos Léxicos | Materiales Divulgativos | Total |
|---|---|---|---|---|
| C.Español | | | | |
| C.Inglés | | | | |

**ANEXO 22 ALGUNAS MEDIDAS EXTERNAS PARA LA EVALUACIÓN DEL AGRUPAMIENTO**

- Entropía (Shannon, 1948) donde m es el número de grupos, nj el tamaño del grupo j, n el número total de objetos agrupados y Ej se calcula según las expresiones en:

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E_j}{n}$$

Entropía de un grupo

$$E_j = -\sum_i p_{ij} \log(p_{ij}) \text{ o } E_j = -\frac{1}{\log m} \sum_{i=1}^{m} p_{ij} \log(p_{ij})$$

donde pij es la probabilidad que un miembro del grupo j pertenezca a la clase i.

Overall F-Measure (OFM)

$$Overall\ F-Measure = \sum_{i=1}^{k} \frac{n_i}{n} \max\{F-Measure(i,j)\}$$

donde k es el número de clases, ni es el tamaño de la clase i, n es el número total de objetos agrupados y F-Measure(i, j) se calcula según la expresión siguiente. Si α=1, entonces OFM.

F-Measure de la clase i respecto al grupo j

$$F-Measure(i,j) = \frac{1}{\alpha\left(1/\Pr(i,j)\right) + (1-\alpha)\left(1/\mathrm{Re}(i,j)\right)}$$

Si α=1 entonces F-Measure (i, j) coincide con precision, si α=0 entonces F-Measure (i,j) concuerda con recall. α=0.5 representa igual peso para precision y recall.

Micro-averaged precision y micro-averaged recall

$$\text{MA - Pr} = \frac{\sum_{i=1}^{k} \alpha_i}{\sum_{i=1}^{k} (\alpha_i + \beta_i)} \text{ y } \text{MA - Re} = \frac{\sum_{i=1}^{k} \alpha_i}{\sum_{i=1}^{k} (\alpha_i + \gamma_i)}$$

donde αi es el número de entes educadamente fijados a la clase i, βi es el número de objetos erróneamente fijados a la clase i y γi es el número de objetos incorrectamente no asignados a la clase i. MA-Pr=MA-Re si cada objeto pertenece a sólo un grupo y la clasificación de referencia también tiene una clasificación única para cada objeto.

Información mutua

$$MI = \sum_{i,j} p_{ij} \cdot \log_2 \frac{p_{ij}}{p_i \cdot p_j}$$

donde pi y pj indican las probabilidades de que un objeto concierna a la clase i y al grupo j, respectivamente, y pij indica la posibilidad de que el objeto pertenezca a la clase i y al grupo j simultáneamente. Esta expresión se normaliza dividiendo por la máxima entropía.

Error del agrupamiento normalizado en el intervalo

$$NCE = \frac{E}{A_t}, \text{donde } A_t = A_m + A_a$$

donde At es el número total de asociaciones que existen en ambas particiones sin eliminar duplicados, donde Am es el número total de asociaciones en la partición de referencia y Aa es el número total de asociaciones en la partición resultado del agrupamiento.

Cluster Recall y Cluster Precisión

$$CR = A_c / A_m \text{ y } CP = A_c / A_a$$

donde Ac=Aa-Ei, simboliza el número total de asociaciones resultantes del agrupamiento.

Rand Statistic

$$R = (a+b)/m$$

Coeficiente de Jaccard

$$J = a/(a+b+c)$$

Índice de Folkes y Mallows

$$FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c}\right)^{1/2}$$

donde a es el número de pares de objetos que conciernen al mismo grupo y a la misma clase, b es el número de aquellos pares que atañen al mismo grupo y a clases diferentes, c es el total de pares que pertenecen a grupos diferentes y a igual clase, d es el número de pares de objetos que pertenecen a grupos y clases disímiles y m=a+b+c+d es el número máximo de todos los pares de objetos (es decir, m=n(n-1)/2 donde n es el número total de objetos).

**ANEXO 23. ALGUNAS MEDIDAS INTERNAS PARA LA EVALUACIÓN DEL AGRUPAMIENTO**

Overall similarity

$$OverallSimilarity(Grupo) = \frac{1}{|Grupo|^2} \sum_{O_i, O_j \in Grupo} \text{distancia}(O_i, O_j)$$

Índices Dun

$$I(C) = \frac{\min_{i \neq j} \{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k} \{\Delta(C_l)\}}$$

donde C={C1, …, Ck} es la asociación de un conjunto de objetos O, δ:C×C→R es una medida de distancia de grupo a grupo y U:C→R es una medida de diámetro del grupo.

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = \max_{x, y \in C_i} d(x, y)$$

donde d:C×C→R es una función que mide la distancia entre los objetos de O.

Una de las propuestas de Bezdek para el cálculo de ( ) jiCC , δ y () iC Δ

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = 2\left( \frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|} \right)$$

donde ci es el centro del grupo Ci.

Índice Davies – Bouldin

$$DB(C) = \frac{1}{k} \cdot \sum_{i=1}^{k} R_i$$

$$R_i = \max_{i \neq j} R_{ij}, \text{ donde } R_{ij} = \frac{(s(C_i) + s(C_j))}{\delta(C_i, C_j)} \text{ y } s(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} |x - c_i|$$

donde C = {C1, …, Ck} representa un agrupamiento de objetos, ci denota el centro del grupo Ci,y s:C→R mide la dispersión dentro del grupo y δ:C×C→R mide la distancia entre grupos. Las medidas Λ y ρ incluyen la colección de objetos como un grafo pesado G=(V,E,w) con el conjunto de nodos V, aristas E y la función de peso w:E→[0,1] donde V simboliza los objetos y w define la similitud entre dos objetos adyacentes.

Medida de conectividad parcial pesada Λ

$$\Lambda(C) = \sum_{i=1}^{k} |C_i| \cdot \lambda_i$$

donde λi designa la conectividad de las aristas pesadas de G(Ci). λ de un grafo G=(V, E, w) es definida como ∑ '},{),( min Evuvuw , donde E' incluye E y G'=(V,

E\E') es no conexo. λ es a la vez designada como la capacidad de un corte mínimo de G.

Medida de densidad esperada ρ

$$\rho(C) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}, \text{ donde } |V|^{\theta} = w(G) \text{ y } w(G) = |V| + \sum_{e \in E} w(e)$$

donde θ se calcula para grafos ponderados según la expresión

$$w(G) = |V|^{\theta} \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)}$$

Modularidad (Modularity)

$$Q = \sum_{i} \left(e_{ii} - a_i^2\right) = \mathrm{Tr}\, \mathbf{e} - \|\mathbf{e}^2\|$$

donde e es una matriz simétrica de orden k cuyo elemento eij es la razón de todas las aristas en el grafo que conectan nodos del grupo i con nodos del grupo j, e indica que la suma de los elementos de la matriz es la traza de la matriz que da la razón de aristas en el grafo que conectan nodos en el mismo grupo.

## ANEXO 24 DISTANCIAS, SIMILITUDES Y DISIMILITUDES MÁS USADAS PARA COMPARAR OBJETOS

Dados los objetos Oi y Oj descritos por k rasgos, donde Oi=(oi1, …, oik) y Oj=(oj1, …, ojk)

- Distancia Euclideana

$$D_{Euclideana}(O_i,O_j)=\sqrt{\sum_{h=1}^{k}(o_{ih}-o_{jh})^2}$$

- Distancia Minkowski

$$D_{Minkowski}(O_i,O_j)=\left(\sum_{h=1}^{k}\left|o_{ih}-o_{jh}\right|^{\gamma}\right)^{\frac{1}{\gamma}} \text{ donde } \gamma \geq 1$$

La distancia Minkowsky es similar a la distancia Manhattan o city-block, y a la distancia Euclideana si se cumple la condición de que γ es 1 y 2, respectivamente (Batchelor 1978). Para los valores de γ ≥ 2, la distancia Minkowsky semeja a Supermum Distancia Euclideana heterogénea (Heterogenous Euclidean – Overlap Metric; HEOM)

$$D_{HEOM}(O_i,O_j)=\sqrt{\sum_{h=1}^{k}d_{local}(o_{ih},o_{jh})^2}\text{, donde}$$

$$d_{local}(o_{ih},o_{jh})=\begin{cases}d_{Overlap}(o_{ih},o_{jh}) & si\ h\ simbólico\\ d_{NormEuclidean}(o_{ih},o_{jh}) & si\ h\ numérico\end{cases}$$

$$d_{Overlap}(o_{ih},o_{jh})=\begin{cases}0, & si\ o_{ih}=o_{jh}\\ 1, & en\ otro\ caso\end{cases} \text{ y } d_{NormEuclidean}(o_{ih},o_{jh})=\frac{\left|o_{ih}-o_{jh}\right|}{\max_h-\min_h}$$

- Distancia Canberra

$$D_{Canberra}(O_i,O_j)=\sum_{h=1}^{k}\frac{\left|o_{ih}-o_{jh}\right|}{\left|o_{ih}+o_{jh}\right|}$$

Correlación de Pearson

$$D_{Pearson}(O_i,O_j)=\frac{\sum_{h=1}^{k}(o_{ih}-\overline{atributo_h})(o_{jh}-\overline{atributo_h})}{\sqrt{\sum_{h=1}^{k}(o_{ih}-\overline{atributo_h})^2\sum_{h=1}^{k}(o_{jh}-\overline{atributo_h})^2}}$$

donde $\overline{atributo_h}$ es el valor promedio que toma el $atributo_h$ en el conjunto de datos.

Coeficiente Dice

$$S_{Dice}(O_i,O_j)=\frac{2\sum_{h=1}^{k}(o_{ih}\cdot o_{jh})}{\sum_{h=1}^{k}o_{ih}^2+\sum_{h=1}^{k}o_{jh}^2}$$

Coeficiente de Jaccard

$$S_{Jaccard}(O_i, O_j) = \frac{\sum_{h=1}^{k} (o_{ih} \cdot o_{jh})}{\sum_{h=1}^{k} o_{ih}^2 + \sum_{h=1}^{k} o_{jh}^2 - \sum_{h=1}^{k} (o_{ih} \cdot o_{jh})}$$

Coeficiente Coseno

$$S_{Coseno}(O_i, O_j) = \frac{\sum_{h=1}^{k} (o_{ih} \cdot o_{jh})}{\sqrt{\sum_{h=1}^{k} o_{ih}^2 \cdot \sum_{h=1}^{k} o_{jh}^2}}$$

**ANEXO 25 REDACCION DE RESUMENES POR EXTRACCIO Y ABSTRACCION**

Cros. Mediante esta indicación se les piden realicen 10 resúmenes utilizando las técnicas de abstracción y otros 10 utilizando solo la técnica de extracción con vistas a realizar la comparación de los resúmenes que ustedes generen contra otros modelos de escritura de texto.

Resultados:

SE REALIZA UN ESTUDIO GEOLOGICO DEL VALLE DE LA LAGUNA (TENERIFE) MEDIANTE CARTOGRAFIA SONDEOS ETC. DE LOS SEDIMENTOS QUE RELLENAN EL VALLE DE LA LAGUNA. SE DIFERENCIAN TRES FORMACIONES A TRAVES DE LOS ANALISIS MINERALOGICOS DE LAS ARCILLAS Y SE RELACIONAN CON LA EVOLUCION CLIMATICA Y SEDIMENTOLOGICA DE LA CUENCA. FINALMENTE SE CORRELACIONAN LOS DATOS DE LOS ENSAYOS GEOTECNICOS CON LOS ANALISIS GRANULOMETRICOS MINERALOGICOS Y LAS ORDENACIONES POSTERIORES DE LA FABRICA.

## ANEXO 26 RESULTADOS DE LA APLICACIÓN DE LA GUIA DE OBSERVACION DE ESTRATEGIAS COGNITIVAS

Luego de la aplicación del procedimiento de observación, validado por las investigaciones de Enders-Niguemeyer (2007) citado por Arco (2008) se procede a explicar los resultados de la observación. Después de observar a los resumidores, proceso en la cual no se estimaron parámetros para valores de Z (Normal), debido a que no se trabaja con muestras, sólo poblaciones completas y luego de haber calculado la confiabilidad individual de las observaciones que se mantuvo al nivel de 0.8 , se logró observar a los resumidores en la actividad práctica. Los resultados después de aplicar la guía de observación son los siguientes:

**PROCESO DE LECTURA**

En este proceso se percibe que sólo el 50 % de los observados refieren leer este apartado del texto y la frecuencia para hacer esta acción cognitiva toma valores de 3 en el conjunto de los estudiados. Esto denota que regularmente se leen los sumarios para confeccionar el resumen. La lectura del resumen del autor reporta valores de observación con valor de 5 para un 90 % en un 90 % de los observados, siendo sólo un 10 % los resumidores que no consultan el resumen del autor siempre y lo hacen con regularidad, es decir con valor 3. La lectura de palabras clave alcanza un valor de 5 y el 100 % de los observados realiza esta estrategia cognitiva siempre. La lectura del texto completo teniendo en cuenta la lectura esquemática del texto fue analizada a partir de sub-elementos de análisis y de ello se obtiene una media que arroja lo siguiente : introducción valor 5 y desarrollada por un 100 % de los resumidores, metodología valor 3 y se reporta como una acción ejecutada por el 80% de los representadores de información, discusión reporta valores de 4 y se realiza la acción por el 90 % de los resumidores, resultados reporta un valor 5 y se consulta este segmento textual por el 100 % de los observados. Como indica la guía la media de los valores de la observación (4) arroja que casi siempre se lee el documento completo y que el 90 % de los estudiados leen casi siempre el texto científico.

**ANALISIS DEL DOCUMENTO**

El 100 % de los observados extrae las oraciones de los párrafos casi siempre (4). El 100 % de los investigados delimita los párrafos siempre (5). Todos los resumidores leen siempre imágenes. Sólo un 40 % de los observados hace acciones en la macroestructura del texto siempre, el resto lo hace casi siempre. Un 25 % de los observados siempre delimita siempre las relaciones anafóricas (5) y el 75 % de ellos regularmente delimita las referidas relaciones. Sólo el 10 % de los analizados desarrolla análisis de elementos gramaticales como verbos, sintagmas verbales y conjugaciones con regularidad, el resto 90 % no lo hace nunca y el valor observado es de 1.

**REPRSENTACION**

En este tipo de apartado se analizaron las estrategias cognitivas y se constató que las acciones metacognitivas que más se desarrollan son las que se declaran en el Anexo 6 (ver Anexo 6) donde están declaradas aquellas estrategias que obtienen una frecuencia de ejecución de un 100 y valores de observación 5 o se aquellas estrategias que se hacen siempre. Los Modelos memorísticos esenciales que se observan en los resumidores son las anotaciones, los subrayados, las enumeraciones de frases y las anotaciones.

Estos pudieron constarse gracias a los resúmenes entregados por los observados. No se observan formas diversas organizar el texto o el conocimiento, generalmente en un 100 % utiliza la forma clásica (objetivos, metodología, resultados y conclusiones).

Es importante destacar que esto es un resumen de un informe que contiene 389 estrategias cognitivas.

**RESUMEN ESTRATEGIAS INTELECTUALES METACOGNITIVAS (METAMODELO PARA LA EXTRACCIÓN Y DESAMBIGUACIÓN DE TEXTOS CIENTÍFICOS)**

**Tecnología:** El análisis está encaminado a buscar el problema que se pretende dar a conocer**.**

- Trata los problemas técnicos de la actividad que se analiza.

**Actividades de control:** el resumidor elabora un plan mental para resumir el documento.

- Plan: Decisión de lo que se va a elaborar lo que va a hacer

**Desarrolla las habilidades cognitivas de inferencia**:

- inferir de las declaraciones implícitas y explícitas del campo del conocimiento.

**Forma de lectura:**

- Leer caracteres incluidos los caracteres formales características (balas, cursiva, tamaño de fuente, etc.) léase: Leer secuencialmente
- Eliminar un tema por escrito dejando un espacio
- Dejar algo de espacio y subrayar:
- Subrayar un pasaje y escribir o anotar.

**Escribir en el Original de los resúmenes de los expertos y explorar**

- Explorar una parte del documento incluye: constar la linealidad del tema que se intenta resumir.
- Mantener un tema de información para uso posterior: almacenar información que sea necesaria para aclarar o mejorar el resumen.

**Comenzar a explorar el documento fuente:**

- Ir por el documento y adquirir información por la forma en las siguientes unidades discursivas: (Objetivos, Metodología, Resultados, Conclusiones y Resultados:

**Seguir el documento original haciendo en primer lugar:**

- Búsqueda de inicio de unidades marcado.
- Localizar los pasajes que haya marcado el resumidor.

**Leer otra vez para refrescar la memoria de nivel superior:**

- Analizar los segmentos de nivel superior del documento.
- Desambiguar usando diversas posturas (conjugación de verbos)
- Declaración de Referencias anafóricas
- Declaración de punto final.

**Comprobar la Unidad Textual:**

- Explorar por unidades, por ejemplo, atendiendo a los párrafos de mayor carga semántica.

**Evaluar la pertinencia de la estructura temática del documento**

- Régimen de Pertinencia: Evaluar como pertinente lo que pertenece al documento o a los esquemas de otros documentos relevantes o sea comparación documental por analogía.

**Pertinencia del Resumen: Considerar en el texto aquellos elementos que sean pertinentes**.

- "Pertinentes-texthint": Utiliza pistas textuales del autor para reconocer las declaraciones pertinentes las unidades: Considera la posibilidad de encontrar al principio y al final de documento unidades pertinentes.
- Construir un plan de acuerdo a estas acciones:
    a. Acción de construir: Producir un resumen paso a paso de acuerdo al tipo de resumen.
    b. tema: Crear un tema frase o seleccionar un tema o frase.
    c. forma-incremento: Formular un pasaje de texto incrementando su contenido.
    d. formulación: Formular un pasaje de texto distribuyendo información (distribución) para ellos es pertinente el uso de una formula estándar o readymade.
    e. Utilice un pasaje de texto "readymade" desde el documento fuente.
    f. Utilizar Imágenes e interpretarlas de forma contextual e incluirlas en el resumen.
    g. reorganizar: Reorganizar un trozo de texto para ajustarlo en un nuevo contexto

## ANEXO 27 ALGUNAS METODOLOGIAS PARA LA CONSTRUCCION DE ONTOLOGÍAS

### Metodología de Cyc

La metodología Cyc (1990) consta de algunos pasos que se describen a continuación

- Codificación manual de conocimiento implícito y explícito extraído de diferentes Fuentes.
- Codificación de conocimiento usando herramientas software.
- Delegación de la mayor parte de la codificación en las herramientas.

### Metodología de Construcción de Ontologías de Uschold y King

Esta metodología fue descrita por Uschold y King en 95, la misma supone algunas generalidades para desarrollar ontologías entre ellas están las siguientes:

(1) identificar el propósito.

(2) Localizar los conceptos e identificar las relaciones entre estos conceptos (3) Reglamentar la ontología.

La ontología resultante debe ser justificada mediante diversos documentos normativos, por lo que posee capacidad para la reutilización. Los pasos que la distinguen como sistema ontológico son los siguientes:

- Identificar propósito
- Capturar la ontología
- Codificación
- Integrar ontologías existentes
- Evaluación
- Documentación

### Metodología de Construcción de Ontologías de Grüninger y Fox

En esta metodología, fue descrita por Grüninger y Fox en 95, lo primero que hay que hacer para aplicar este método determinar las aplicaciones donde se insertará la ontología. En autores como Fernández (2003) se expresa que se usa un conjunto de preguntas en lenguaje natural, llamadas cuestiones de competencia, para determinar el ámbito de la ontología. Estas interrogantes son utilizadas para formular y sistematizar los megaconceptos principales, así como su aparato paradigmático) propiedades, relaciones y axiomas) Esta

metodología se usó para construir la ontología TOVE y obliga al uso de los siguientes pasos para la construcción de ontología:

- Campos de Aplicación.
- Características informales de competencia
- Terminología del ámbito.
- Cuestiones formales de competencia
- Axiomas formales
- Teoremas de completitud

**Metodología KACTUS**

En esta metodología diseñada por Bernaras en 1996 se sustenta en el desarrollo de bases de conocimiento para una aplicación determinada. Esta metodología se ha usado para erigir una ontología con vistas a calificar fallos, y se une a dos o más aplicaciones. Los pasos que recomienda para su desarrollo son los siguientes:

- Especificación de la aplicación
- Diseño preliminar basado en categorías ontológicas top-level relevantes
- Refinamiento y estructuración de la ontología.

**METHONTOLOGY**

Methontology es una metodología para construir ontologías a partir de la reusabilidad de sus estructuras o a partiendo del análisis de un dominio propio. Para trabajar con esta metodología es necesario: (1) individualización del proceso de construcción de la Ontología que incluye las siguientes actividades Ñ(evaluación, gestión de configuración, conceptualización, integración, implementación, etc.); (2) un ciclo de vida sustentado en simulaciones pequeñas o prototipos de ontología evolucionados; y (3) la metodología de acción que incluye los siguientes pasos:

- Especificación
- Conceptualización
- Formalización
- Implementación
- Mantenimiento

**Metodología SENSUS**

La metodología basada en Sensus está centrada en una visión orientada a top-down para generar ontologías específicas de una comunidad epistémica a partir de grandes ontologías. El proceder de la ontología consiste en la identificación de un conjunto de términos base cuyo valor está probado en un dominio específico. Tales términos se enlazan manualmente a una ontología de amplia cobertura. El proceso de la construcción de la ontología se realiza con el concurso de los usuarios, pues son los que seleccionan vocablos relevantes para referir el dominio y delimitar la ontología Sensus. Esta metodología cuenta con los siguientes pasos:

- Tomar una serie de términos como términos base.
- Unir lo términos manualmente.
- Determinar todos los conceptos imprescindibles para construir el sustento del sistema.
- Añadir nuevos términos relevantes del dominio.

**Metodología On-To-Knowledge**

El proyecto OTK de Staab 2001, desarrolla ontologías con el objeto de hacer óptima la gestión del conocimiento en organizaciones de diversas dimensiones. El procedimiento suministra herramientas que propician la introducción de conceptos e instrumentos para gestionar y manejar conocimiento empresarial, pues facilitan la construcción de estrategias los proveedores y buscadores de conocimiento a presentar éste de forma eficiente y efectiva. Esta metodología obliga que la organización que la aplica que la desarrolla tenga un alto desempeño en la gestión del conocimiento. Los siguientes pasos son recomendados por esta metodología:

- Estudio de viabilidad
- Comienzo
- Refinamiento
- Evaluación

**ANEXO 28 RESMEN REGLAS PUERTOT 1 A**

**Introducción**

*El impacto de la micromedición en un sistema de distribución de agua potable tiene como objeto determinar la factibilidad de instalar micromedidores domiciliarios en zonas de la población donde no existen aún los aparatos.*

Objetivos y Metodología

*En este artículo se presenta el impacto de la micromedición en la ciudad de Guaymas. Son., en donde primero se instalaron, durante dos meses, micromedidores ocultos a tres grupos de usuarios pertenecientes a las clases socioeconómicas baja, media y alta; posteriormente a los mismos usuarios se les instaló el aparato visible.*

**Resultados**

Según los resultados, la clase socioeconómica media es donde se registró el mayor impacto, reduciendo sus consumos hasta un 48%.

**Discusión**

Asimismo, el 68% de los usuarios seleccionados consumió menos agua con micromedidor; en el 32% restante no hubo ningún efecto.

## ANEXO 29 RESUMEN PUERTOT 1 B

### Introducción

*El subsistema Sur de Sierra de Gádor Campo de Dalías está integrado por un conjunto de acuíferos relacionados entre sí y con el mar, afectados por tendencias indeseables de diferente tipo que responden a la utilización que ha venido soportando en los últimos cuarenta años. Su compleja estructura e intenso bombeo hacen necesario, de forma más acusada que en otros medios hidrogeológicos, el conocimiento actualizado de su funcionamiento y procesos asociados de contaminación particularizados por acuíferos y subacuíferos. Esta información resulta imprescindible para llevar a cabo una*
*gestión equilibrada de aprovechamiento sostenible del subsistema, hasta ahora carente de una gestión real.*

### Objetivos

*Está previsto realizar una desaladora para sustitución de parte del bombeo de este subsistema*.

### Resultados

*Para cuando estos nuevos recursos estén disponibles, deberían estar seleccionadas las actuaciones de gestión conducentes a la recuperación y protección de estos acuíferos, lo que habría de realizarse asumiendo su complejidad, con el fin de gestionar los recursos hídricos desalados y subterráneos con racionalidad, es decir, usando el importante conocimiento adquirido ya sobre el funcionamiento hidrogeológico de estos acuíferos, lo que permitiría discriminar los procesos negativos que requieren mayor corrección, para una gestión sostenible, orientando el diseño y aplicación de las medidas más adecuadas para evitar su progresión y proteger a los acuíferos de sus efectos.*

### Discusión

De estos efectos negativos se exponen aquí algunos ejemplos.

## ANEXO 30 RESUMEN REGLAS PUERTOT 2 A

**Objetivo**

El presente estudio analiza el impacto que el uso de sistemas de riego por aspersión tiene sobre la productividad del cultivo de alfalfa en esta región. Metodología comparando dos sistemas de riego (pivote central y power roll con intervalos cortos y moderados de aplicación), con el riego superficial.

Resultados

Bajo riego por aspersión se advirtió mayor población y cobertura de planta, así como mayor rendimiento y longevidad del cultivo; en el área de macollo no hubo diferencia.

Conclusiones

El análisis económico del estudio mostró que el uso de sistemas de riego pivote central y power roll alcanzan tasas de retorno marginal mayores a 244%, es decir, se concluye que su adopción en el cultivo de alfalfa, resulta más importante desde el punto de vista de los rendimientos y reditucbilidad que bajo riego superficial.

## ANEXO 31 RESUMEN REGLAS PURTOT 2 b

### Introducción

Frente a la necesidad de lograr un incremento en la eficiencia en el uso del agua de riego, en México se han tomado diversas medidas, una de las cuales es la instauración de la dotación volumétrica a los usuarios de los sistemas de riego, lo que en conjunto con otras acciones permitirá un mejor control del agua empleada.

Adicionalmente, pueden calibrarse analíticamente, lo que supone importantes ahorros en calibraciones en campo. No obstante, el diseño de estos aforadores y su calibración analítica no son una tarea fácil; la calibración supone la solución de ecuaciones diferenciales del flujo y de la capa límite, y deben probarse varias alternativas antes de obtener un diseño satisfactorio.

### Metodología

Para evitar las dificultades citadas, se ha desarrollado un programa de computadora, que permite estudiar opciones de diferentes geometrías, analizar los efectos sobre el flujo en el canal, y calcular y dibujar curvas de calibración, entre otras opciones.

### Conclusiones

En este artículo se presentan la metodología, técnicas, criterios y ecuaciones empleadas en el programa, se describe su operación y se presentan los resultados de su validación en laboratorio y en campo

## ANEXO 32 RESUMEN BASELINE

Varias áreas en Oregón, EUA, son susceptibles de ser contaminadas por fuentes no puntuales de contaminación que provienen de cultivos bajo riego. Un uso eficiente y un reúso económico del agua en riego es la clave para proteger la calidad y la cantidad de las aguas subterráneas superficiales. La salinidad del suelo no es un gran problema en estas regiones. La agricultura ha estado preocupada casi exclusivamente del uso efectivo de las entradas de energía para obtener producciones máximas. Ahora este enfoque ha cambiado hacia el manejo de los recursos para la preservación de la calidad del agua y de la tierra. Varios incentivos y normatividades están reforzando este cambio en Oregón. El manejo de las estrategias para la protección de la calidad del agua a través del riego eficiente incluye el tiempo y proporción de la aplicación de químicos, métodos y técnicas de riego más eficientes, y el reúso del agua para riego. El agua subterránea con contenido de nitratos está siendo usada para riego, dado que tiene valor de fertilizante y de agua.

## ANEXO 33 RESUMEN MICROSOFT-WORD

**Introducción**

Después de una serie de trabajos de campo combinados con estudios de laboratorio se ha llegado a la conclusión de que las magnesitas del Yacimiento de Asturreta (Navarra) se han formado por Procesos de Metasomatismo que produjeron un reemplazamiento de la roca original (Dolomita) por magnesita debido a la acción de soluciones hidrotermales ricas en $C0_2$ y a temperaturas relativamente bajas no llegándose a alcanzar los 350c

**Metodología**

El proceso debió producirse en un ambiente de relativa tranquilidad que permitió el que se formase grandes cristales de magnesita dispuestas en empalizada y que se respetasen las estructuras sedimentarias que se observan posteriormente tuvo lugar una redolomitación

**Objetivos**

**¿**

**Resultados**

**¿**

**Conclusiones**

**¿**

**Resumen**

Después de una serie de trabajos de campo combinados con estudios de laboratorio se ha llegado a la conclusión de que las magnesitas del Yacimiento de Asturreta (Navarra) se han formado por Procesos de Metasomatismo que produjeron un reemplazamiento de la roca original (Dolomita) por magnesita debido a la acción de soluciones hidrotermales ricas en $C0_2$ y a temperaturas relativamente bajas no llegándose a alcanzar los 350c el proceso debió producirse en un ambiente de relativa tranquilidad que permitió el que se formase grandes cristales de magnesita dispuestas en empalizada y que se respetasen las estructuras sedimentarias que se observan posteriormente tuvo lugar una redolomitación

## ANEXO 34 COMPETENCIA DE LOS EXPERTOS

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | A | 1 | 1 |
| 2 | 0.5 | 0.5 | 0.5 | A | 0.5 | 0.5 |
| 3 | 0.46 | 0.5 | 0.43 | M | 0.3 | 0.3 |
| 4 | 0.48 | 0.5 | 0.46 | M | 0.5 | 0.5 |
| 5 | 0.45 | 0.5 | 0.4 | M | 0.3 | 0.3 |
| 6 | 0.46 | 0.5 | 0.43 | M | 0.5 | 0.3 |
| 7 | 0.46 | 0.5 | 0.43 | M | 0.5 | 0.5 |
| 8 | 0.46 | 0.5 | 0.43 | M | 0.3 | 0.3 |
| 9 | 0.46 | 0.5 | 0.43 | M | 0.5 | 0.5 |
| 10 | 0.46 | 0.5 | 0.43 | M | 0.5 | 0.5 |
| 11 | 0.46 | 0.5 | 0.43 | M | 0.5 | 0.5 |

**ANEXO 35 EVALUACION POR EXPERTOS DEL MODELO TEXMINER (Resultados)**

Sr(a): Me llamo Amed Abel Leiva Mederos, estudiante de Doctorado de la Universidad de Granada, mediante este texto le invito a evaluar el modelo de construcción automática de resúmenes que propongo en mi tesis, por ser usted uno de los expertos con elevada calificación en el tema. La lectura del modelo solo exigirá de unos 15 minutos y cumplimentar la encuesta también. Le agradecería que brindara sus observaciones y valoraciones sobre la propuesta.

**Nombre**: TEXMINER. Modelo para la extracción y desambiguación de textos científicos.

El modelo que se presenta tiene como ***objetivo general***: Construir resúmenes automáticos a partir de textos científicos.

Dentro de las ***características generales*** del modelo son:

13. Estudio de Necesidades.
14. Análisis Manual del Corpus Textual.
15. Creación de la ontología.
16. Extracción del Texto mediante Agentes Cognitivos.
17. Modelación de sistema de búsqueda y recuperación de información.
18. Representación de la Información

Teniendo en cuenta estas particularidades, se establecieron los elementos necesarios para la construcción del modelo, considerando como dispositivos imprescindibles para lograr su efectividad a los procesos que aquí se describen. La teoría aquí planteada emana de de los diversos instrumentos y métodos desarrollados en la investigación:

- Examinar los referentes teóricos metodológicos que han afectado al resumen automático y su desarrollo, haciendo énfasis en las preposiciones que han aparecido en Ciencias de la Computación, Ciencias de la Información, la Lingüística, la Semiótica y la Cibersemiótica.
- Estudio de las estrategias cognitivas de 12 resumidores para establecer regularidades en el proceso de resumir.
- Estudiar el dominio textual a través del análisis de discurso para establecer ejes semánticos.

- Construcción de reglas textuales a partir de las regularidades establecidas en los textos e implementar dichas reglas en agentes o robots que simulen estas regularidades.
- Evaluar la calidad de los textos a través de herramientas de análisis estadístico y con el criterio de los especialistas en lingüística.
- Utilizar herramientas de visualización que permitan la observación del contexto de cada concepto.
- Permitir la evaluación de los textos resultantes y el acceso a los textos originales, dotando al sistema de una real y lógica representación textual.
- Construir un sistema de extracción de textos que valide en la práctica el modelo.
- Evaluar el sistema teniendo en cuenta criterios de usabilidad.

A partir de las propuestas teóricas y procedimentales desarrolladas se determinaron los siguientes componentes para el modelo:

***Profesionales encargados del proceso***.

Este rol es importante en el modelo, ya que un sistema que trabaje con texto y con técnicas de inteligencia artificial necesita de un personal heterogéneo, por esta razón para ello se escoge:

4. **Bibliotecario:** Con una preparación que rebase los principios de la catalogación y la representación ordinaria, asumiendo formas de descripción estructurada, entiéndase por esto sistemas de marcado xml, estructuración semántica de contenidos, etc.
5. **Ingeniero Informático:** Uno de los errores en los procesos de desarrollo de aplicaciones para el tratamiento de información ha sido dejar en el terreno de los informáticos múltiples aplicaciones. En este caso el papel del informático es programar las operaciones cognitivas que detecten los bibliotecarios a través de las actualizaciones del modelo.
6. **Usuarios:** Juegan un papel fundamental en el desarrollo del sistema y del modelo, sus estrategias cognitivas son la base de los procesos que se automatizan y hacia ellos va dirigido el producto final: el resumen automático. Todos los procesos de representación visual son desarrollados a partir de las necesidades de la comunidad de usuarios.

Estos roles constituyen una relación indispensable que no distingue de niveles en el modelo, demostrando la indisoluble unión entre los que revisan los contenidos, los que los distribuyen y quienes los consumen o usan.

***II. Recursos***

Por esta categoría se entiende todos los medios, los métodos y procedimientos que tributen al modelo. Esto propició reconocer elementos que destacan la relación biblioteca, organización y usuarios de la información.

Para lograr esta relación se necesitan, los planes de desarrollo de las organizaciones en los que se reflejen diversos datos como: , sus publicaciones, sus profesionales, así como el efectivo registro de necesidades de usuarios, tanto individuales como colectivos, además también se analiza el desarrollo de competencias informacionales tanto para usuarios como para todo el personal con roles en el modelo. Otros recursos que demanda el modelo son los tecnológicos, entre los que se encuentran máquinas y redes de computadoras.

***III. Interacción entre los componentes***

Todos los componente están relacionados, tanto a nivel procedimental (recogida de información) como técnico (unión correcta de las aplicaciones), cognitivo (estructuración del conocimiento). TEXMINER es el resultado de la interacción de los últimos paradigmas de la Representación de la información y el Conocimiento, específicamente el sociocognitivismo, que obliga a determinar las especificidades de la realidad representacional de cada usuario para utilizarla en los modelos de tratamiento de información, alejando los presupuestos de la organización y el resumen del paradigma físico, sustentado en la lingüística clásica para entrar en el terreno de la representación conceptual, más ligada a las vivencias de las comunidades de práctica. La interdependencia entre los componentes puede ser valorada a partir de los siguientes principios:

- La calidad de los resúmenes que ofrece a los usuarios.
- La capacidad para visualizar conceptos a partir de estrategias cognitivas.
- Descripción y Catalogación de Recursos a partir de modelos de metadatos más flexibles

- Búsqueda de información a partir de la semántica que declaren los conceptos
- La generación de redes sociales a partir de dominios específicos, lo que permitirá la comunicación interprofesional.

**Pasos del Modelo**

El gráfico declara las etapas del modelo, las mismas están enumeradas dando a entender la estructuración holística de todas las etapas, de forma que se genera un ciclo donde después de la evaluación del resumen se declaren nuevas necesidades de información que estén sustentadas en nuevas estrategias cognitivas. Hay por ende un modelo holístico para el desarrollo del resumen declarando aquellos elementos imprescindibles que desde la teoría de la Ciencia de la Información se ha exigido para estos materiales.

Cuestionario

De los principios del Modelo TEXMINER exprese su evaluación a través de los siguientes valores

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | 11 | 0 | 0 | 0 | 0 | 0 |
| 2 | 10 | 2 | 0 | 0 | 0 | 0 |
| 3 | 10 | 1 | 0 | 0 | 0 | 0 |
| 4 | 9 | 3 | 0 | 0 | 0 | 0 |
| 5 | 10 | 2 | 0 | 0 | 0 | 0 |
| 6 | 11 | 0 | 0 | 0 | 0 | 0 |

De los pasos declarados en la concepción teórica del modelo TEXMINER exprese su valoración a través de los siguientes indicadores.

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | 11 | 0 | 0 | 0 | 0 | 0 |
| 2 | 9 | 2 | 0 | 0 | 0 | 0 |
| 3 | 11 | 0 | 0 | 0 | 0 | 0 |
| 4 | 8 | 2 | 1 | 0 | 0 | 0 |
| 5 | 11 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | 10 | 1 | 1 | 0 | 0 | 0 |
| 7 | 9 | 2 | 1 | 0 | 0 | 0 |
| 8 | 11 | 0 | 0 | 0 | 0 | 0 |

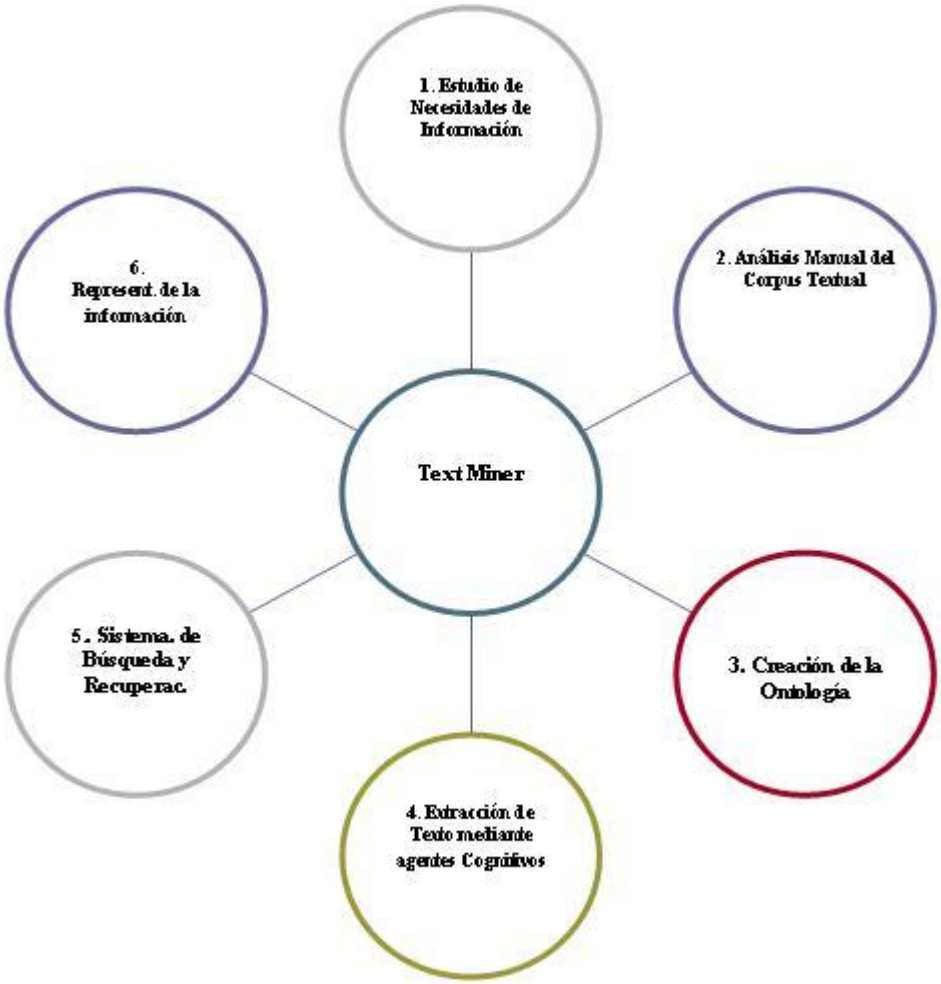
Realice cualquier valoración que considere sobre el modelo propuesto

Sobre la integración de los componentes en el modelo responda:

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | 11 | | | | | |
| 2 | 10 | 1 | | | | |
| 3 | 11 | | | | | |
| 4 | 11 | | | | | |
| 5 | 11 | | | | | |

Sobre los profesionales integrados al modelo responda:

| Pasos | Muy Importante | Bastante Importante | Importante | Poco Importante | No Importante | Total |
|---|---|---|---|---|---|---|
| 1 | 11 | | | | | |
| 2 | 11 | | | | | |
| 3 | 11 | | | | | |


1. Estudio de Necesidades de Información
2. Análisis Manual del Corpus Textual
3. Creación de la Ontología
4. Extracción de Texto mediante agentes Cognitivos
5. Sistema. de Búsqueda y Recuperac.
6. Represent. de la información
Text Miner

## ANEXO 36 DTD DEL SISTEMA

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dtc="dtcelements.rdf#">
        <rdf:Description id="001">
                <dc:title>ESTUDIO DE CORRIENTES Y DISPERSIÃ“N EN EL
PUERTO DE BARCELONA</dc:title>
                <dc:creator>
                        <rdf:Bag>
                                <rdf:li>M. DÃ-ez</rdf:li>
                                <dtc:afiliacion>Departamento de FÃ-sica Aplicada.
Universidad de CataluÃ±a</dtc:afiliacion>
                                <dtc:rdfli>J.M.Redondo</dtc:rdfli>
                                <dtc:afiliacion>Autoridad Portuaria de Barcelona. APB
Medio Ambiente</dtc:afiliacion>
                                <dtc:rdfli>J.Vila</dtc:rdfli>
                                <dtc:afiliacion>Autoridad Portuaria de Barcelona.
Universidad de CataluÃ±a</dtc:afiliacion>
                         </rdf:Bag>
 </dc:creator>
                <dtc:introduccion>
                        <dtc:C>
                                <dtc:Ora>
                                        <dtc:C>
                                                <dtc:elab_n>
                                                        <dtc:E>
                                                                <dtc:conector>
                                                                        <dtc:text>Uno
</dtc:text>
                                                                </dtc:conector>
                                                        </dtc:E>
                                                        <dtc:E>
                                                                <dtc:cuerp>
                                                                        <dtc:text> de los
problemas medioambientales de los puertos</dtc:text>
                                                                </dtc:cuerp>
```

```
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:fv>
                        <dtc:text> es </dtc:text>
                                                    </dtc:fv>
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:procesos>
                                                            <dtc:text> el vertido
de contaminantes </dtc:text>
                                                        </dtc:procesos>
                                                    </dtc:E>
                                                </dtc:elab_n>
                                            </dtc:C>
                                            <dtc:C>
                                                <dtc:elab_s>
            <dtc:E>
            <dtc:text> en especial de hidrocarburos </dtc:text>
            </dtc:E>
        </dtc:elab_s>
                                            </dtc:C>
                                        </dtc:Ora>
                                    </dtc:C>
                                    <dtc:C>
                                        <dtc:Ora>
                                            <dtc:C>
                                                <dtc:just_n>
                                                    <dtc:E>
                                                        <dtc:conector>
                                                            <dtc:text> Dado
</dtc:text>
                                                        </dtc:conector>
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:sub>
                                                            <dtc:text> que
</dtc:text>
```

```
                                                    </dtc:sub>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:procesos>
                                                        <dtc:text> la
dispersiÃ³n </dtc:text>
                                                    </dtc:procesos>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:fv>
                        <dtc:text> es </dtc:text>
                    </dtc:fv>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:conector>
            <dtc:text> un </dtc:text>
        </dtc:conector>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:procesos>
                                                        <dtc:text>
mecanismo </dtc:text>
                                                    </dtc:procesos>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:sub>
                                                        <dtc:text> que
</dtc:text>
                                                    </dtc:sub>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:fv>
                         <dtc:text> hace crecer </dtc:text>
                    </dtc:fv>
                                                </dtc:E>
                                                <dtc:E>
                                                    <dtc:cuerp>
```

<dtc:text>
exponencialmente el tamaÃ±o de las manchas </dtc:text>
</dtc:cuerp>
</dtc:E>
<dtc:C>
<dtc:just_s>
<dtc:E>
<dtc:sub>

<dtc:text> y que </dtc:text>

</dtc:sub>
</dtc:E>
<dtc:E>
<dtc:cuerp>

<dtc:text> el coste de su limpieza y daÃ±os tambiÃ©n es funciÃ³n del tamaÃ±o
</dtc:text>
</dtc:cuerp>
</dtc:E>
<dtc:E>
<dtc:fv>

<dtc:text> es </dtc:text>

</dtc:fv>
</dtc:E>
<dtc:E>
<dtc:cuerp>

<dtc:text> imprescindible un buen conocimiento del fenÃ³meno </dtc:text>
</dtc:cuerp>
</dtc:E>
<dtc:E>
<dtc:sub>

<dtc:text> que </dtc:text>

</dtc:sub>
</dtc:E>

```
                                                    <dtc:E>
                                                        <dtc:cuerp>

        <dtc:text> junto con un buen sistema de alerta y acciÃ³n </dtc:text>
                                                        </dtc:cuerp>
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:fv>
permita </dtc:fv>
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:cuerp>

        <dtc:text> una rÃ¡pida intervenciÃ³n </dtc:text>
                                                        </dtc:cuerp>
                                                    </dtc:E>
                                                </dtc:just_s>
                                            </dtc:C>
                                        </dtc:just_n>
                                    </dtc:C>
                                </dtc:Ora>
                            </dtc:C>
                        </dtc:introduccion>
                        <dtc:Desarrollo>
                            <dtc:C>
                                <dtc:Ora>
                                    <dtc:C>
                                        <dtc:elab_n>
                                            <dtc:E>
                                                <dtc:conector>
                                                    <dtc:text> En
</dtc:text>
                                                </dtc:conector>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:lug>
```

<dtc:text> el Puerto de Barcelona </dtc:text>
</dtc:lug>
</dtc:E>
<dtc:E>
<dtc:fv>
<dtc:text> se estÃ¡n </dtc:text>
</dtc:fv>
</dtc:E>
<dtc:E>
<dtc:cuerp>
<dtc:text> llevando a cabo una serie de experimentos </dtc:text>
</dtc:cuerp>
</dtc:E>
<dtc:C>
<dtc:just_s>
<dtc:E>
<dtc:text> con el fin de </dtc:text>
</dtc:E>
<dtc:E>

<dtc:fv>

<dtc:text> estudiar </dtc:text>

</dtc:fv>
</dtc:E>
<dtc:E>
<dtc:text> el campo de corrientes y la capacidad dispersiva del medio </dtc:text>
</dtc:E>
</dtc:just_s>
</dtc:C>
</dtc:elab_n>
</dtc:C>

```
                    </dtc:Ora>
                </dtc:C>
                <dtc:C>
                    <dtc:Ora>
                        <dtc:C>
                            <dtc:elab_n>
                                <dtc:E>
                                    <dtc:fv>
                <dtc:text> EstÃ¡n previstas </dtc:text>
                </dtc:fv>
                                <dtc:E>
                                <dtc:E>
                                    <dtc:cuerp>
                                        <dtc:text>
campaÃ±as para cada </dtc:text>
                                    </dtc:cuerp>
                                <dtc:E>
                                <dtc:E>
                                    <dtc:sustadj>
                                        <dtc:text> estaciÃ³n
climÃ¡tica</dtc:text>
                                    </dtc:sustadj>
                                <dtc:E>
                            </dtc:elab_n>
                            <dtc:C>
                                <dtc:elab_s>
                                    <dtc:E>
                                        <dtc:conector>
                                            <dtc:text>
asÃ- como </dtc:text>
                                        </dtc:conector>
                                    </dtc:E>
                                    <dtc:E>
                                        <dtc:sustadj>
                                            <dtc:text>
campaÃ±as especiales </dtc:text>
                                        </dtc:sustadj>
```

```
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:conector>
                                                                            <dtc:text>
para </dtc:text>
                                                                        </dtc:conector>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:fv>
                                    <dtc:text> analizar </dtc:text>
                                </dtc:fv>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:fenomenos>
                                                                            <dtc:text>
fenÃ³menos hidrodinÃ¡micos </dtc:text>
                                                                        </dtc:fenomenos>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:conector>
                                    <dtc:text> como la </dtc:text>
                                </dtc:conector>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:fenomeno>
                                                                            <dtc:text>
difracciÃ³n o estratificaciÃ³n </dtc:text>
                                                                        </dtc:fenomeno>
                                                                    </dtc:E>
                                                                </dtc:elab_s>
                                                            </dtc:C>
                                                        </dtc:C>
                                                    </dtc:Ora>
                                                </dtc:C>
                                                <dtc:C>
                                                    <dtc:Ora>
                                                        <dtc:C>
```

```
                                                    <dtc:elab_n>
                                                        <dtc:E>
                                                            <dtc:conector>
                                                                <dtc:text> La
</dtc:text>
                                                            </dtc:conector>
                                                        </dtc:E>
                                                        <dtc:E>
                                                            <dtc:cuerp>
                                                                <dtc:text> duraciÃ³n
de cada campaÃ±a </dtc:text>
                                                            </dtc:cuerp>
                                                            <dtc:E>
                                                                <dtc:fv>
                                <dtc:text> es </dtc:text>
                                </dtc:fv>
                                                            </dtc:E>
                                                            <dtc:E>
                                                                <dtc:cuerp>
                            <dtc:text> de dos semanas </dtc:text>
                            </dtc:cuerp>
                                                            </dtc:E>
                                                        </dtc:E>
                                                        <dtc:C>
                                                            <dtc:just_n>
                            <dtc:E>
                            <dtc:text> debido al gran dominio de trabajo </dtc:text>
                              </dtc:E>
                                  </dtc:just_n>
                                                        </dtc:C>
                                                    </dtc:elab_n>
                                                </dtc:C>
                                            </dtc:Ora>
                                        </dtc:C>
                                        <dtc:C>
                                            <dtc:Ora>
                                                <dtc:C>
```

```
                                        <dtc:elab_n>
                                            <dtc:E>
                                                <dtc:conector>
                <dtc:text> Los </dtc:text>
                    </dtc:conector>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:procesos>
                                                    <dtc:text>
experimentos </dtc:text>
                                                </dtc:procesos>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:fv>
                    <dtc:text> consisten </dtc:text>
                  </dtc:fv>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:conector>
                                                    <dtc:text> en el
</dtc:text>
                                                </dtc:conector>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:procesos>
                                                    <dtc:text> vertido de
trazadores lagrangianos </dtc:text>
                                                </dtc:procesos>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:conector>
                                                    <dtc:text> y la
</dtc:text>
                                                </dtc:conector>
                                            </dtc:E>
                                            <dtc:E>
            <dtc:text> filmaciÃ³nde su comportamiento </dtc:text>
```

```
                </dtc:E>
                                        </dtc:elab_n>
                                    </dtc:C>
                                </dtc:Ora>
                            </dtc:C>
                            <dtc:C>
                                <dtc:Ora>
                                    <dtc:C>
                                        <dtc:elab_s>
                                            <dtc:E>
                                                <dtc:conector>
                                                    <dtc:text> Asimismo
</dtc:text>
                                                </dtc:conector>
                                            </dtc:E>
                                            <dtc:E>
                                                <dtc:cuerp>
                                                    <dtc:text> con un
GPS </dtc:text>
                    </dtc:cuerp>
                </dtc:E>
                <dtc:E>
                                                 <dtc:fv>
                        <dtc:text> se registran </dtc:text>
                    </dtc:fv>
                                                </dtc:E>
                                                <dtc:E>

        <dtc:procesos>
                <dtc:text> los puntos de vertido y recogida </dtc:text>
            </dtc:procesos>
                                                        </dtc:E>
                                                        <dtc:E>

        <dtc:fv>
                            <dtc:text> tomÃ¡ndose </dtc:text>
                        </dtc:fv>
```

```
                                                                    </dtc:E>
                                                                    <dtc:E>
                <dtc:text> como </dtc:text>
              </dtc:E>
               <dtc:E>
                  <dtc:text> puntos fiduciales en la restituciÃ³n digital de las imÃ¡genes
</dtc:text>
                                                                        </dtc:E>
                                                                </dtc:elab_s>
                                                            </dtc:C>
                                                        </dtc:Ora>
                                                    </dtc:C>
                                                    <dtc:C>
                                                        <dtc:Ora>
                                                            <dtc:C>
                                                                <dtc:elab_n>
                                                                    <dtc:E>
                                                                        <dtc:conector>
                                                                            <dtc:text> Los
</dtc:text>
                                                                        </dtc:conector>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:instrumentos>
                                                                            <dtc:text> razadores
</dtc:text>
                                                                        </dtc:instrumentos>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:fv>
                                <dtc:text> son </dtc:text>
                                                                    </dtc:fv>
                                                                    </dtc:E>
                                                                    <dtc:E>
                                                                        <dtc:cuerp>
                                <dtc:text> de dos tipos lagrangianos para </dtc:text>
                                                                    </dtc:cuerp>
```

```
                                                    </dtc:E>
                                                    <dtc:E>
                                                            <dtc:fv>
                        <dtc:text> medir </dtc:text>
                       </dtc:fv>
                                                    </dtc:E>
                                                    <dtc:E>
                                                            <dtc:instrumentos>
                                                                    <dtc:text> corrientes
flotadores </dtc:text>
                                                            </dtc:instrumentos>
                                                    </dtc:E>
                                                    <dtc:E>
                                                            <dtc:cuerp>
                                                                    <dtc:text>
adecuadamente lastrados </dtc:text>
                                                            </dtc:cuerp>
                                                    </dtc:E>
                                                    <dtc:C>
                                                            <dtc:just_n>
                <dtc:E>
                                                                                    <dtc:fv>
                          <dtc:text> para evitar </dtc:text>
                        </dtc:fv>
                 </dtc:E>
                                                                                    <dtc:E>

        <dtc:procesos>

        <dtc:text> la interferencia del viento, y manchas de leche con </dtc:text>

        </dtc:procesos>

        <dtc:fq>

        <dtc:text>  fluoresceÃ-na </dtc:text>
```

```
</dtc:fq>
                                                                    </dtc:E>
                                                                <dtc:E>
                                                                    <dtc:cuerp>

<dtc:text> por su elevado contraste y persistencia </dtc:text>
                                                                    </dtc:cuerp>
                                                                </dtc:E>
                                                            </dtc:just_n>
                                                        </dtc:C>
                                                    </dtc:elab_n>
                                                </dtc:C>
                                            </dtc:Ora>
                                        </dtc:C>
                                    </dtc:Desarrollo>
                                    <dtc:Resultados>
                                        <dtc:C>
                                            <dtc:Ora>
                                                <dtc:E>
                                                    <dtc:conector>
                                                        <dtc:text> A </dtc:text>
                                                    </dtc:conector>
                                                </dtc:E>
                                                <dtc:C>
                                                    <dtc:elab_n>
                                        <dtc:E>
                                        <dtc:text> travÃ©s del anÃ¡lisis digital de imÃ¡genes </dtc:text>
                                        </dtc:E>
                                        <dtc:E>
                                                                    <dtc:fv>
                                            <dtc:text> se obtienen </dtc:text>
                                        </dtc:fv>
                                                            </dtc:E>
                                                        <dtc:E>
                                                            <dtc:cuerp>
```

<dtc:text> las trayectorias, direcciones y velocidades de la corriente y los coeficientes de dispersiÃ³n </dtc:text>

</dtc:cuerp>

</dtc:E>

</dtc:elab_n>

</dtc:C>

</dtc:Ora>

</dtc:C>

<dtc:C>

<dtc:Ora>

<dtc:C>

<dtc:lista_n>

<dtc:E>

<dtc:text> Por otro lado se analizan los agentes impulsores como el viento, marea, oleaje, corrientes, etc </dtc:text>

</dtc:E>

</dtc:lista_n>

</dtc:C>

</dtc:Ora>

</dtc:C>

<dtc:C>

<dtc:Ora>

<dtc:C>

<dtc:res_n>

<dtc:E>

<dtc:conector>

<dtc:text> Los </dtc:text>

</dtc:conector>

</dtc:E>

<dtc:E>

<dtc:cuerp>

<dtc:text> primeros resultados </dtc:text>

</dtc:cuerp>

</dtc:E>

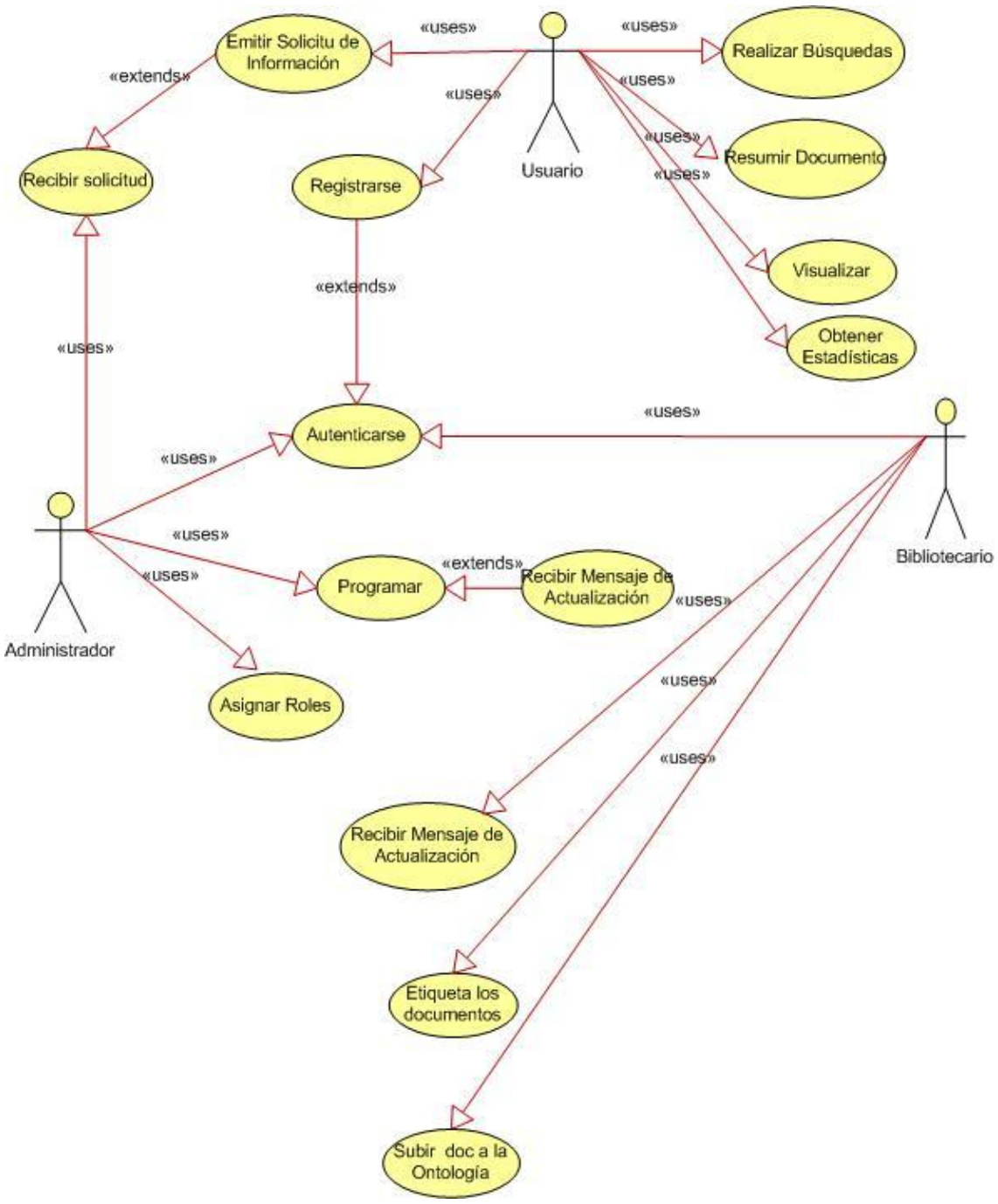
<dtc:E>

<dtc:fv>

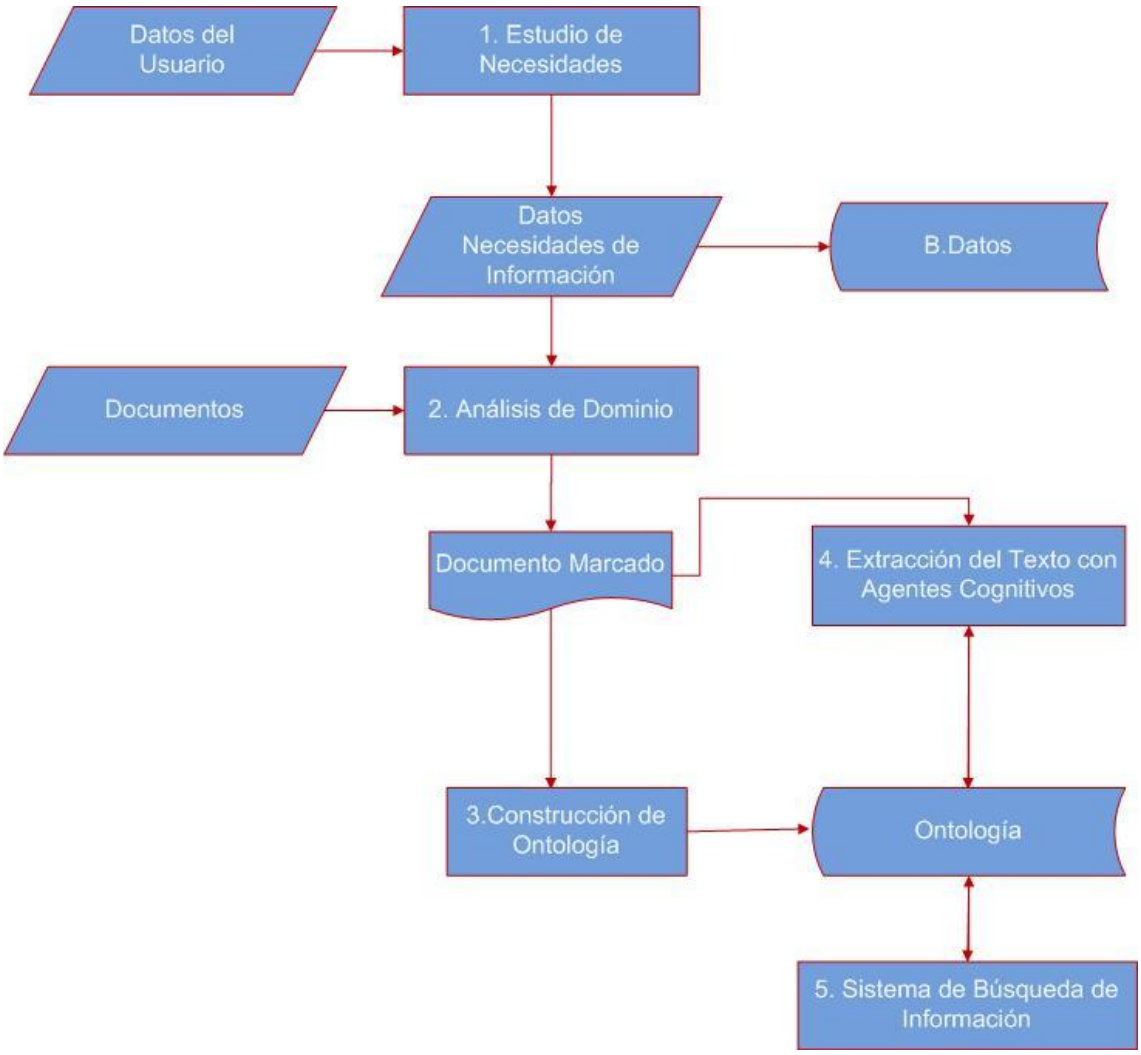
<dtc:text> muestran </dtc:text>
</dtc:fv>
</dtc:E>
<dtc:E>
<dtc:sub>
<dtc:text> que la influencia de la marea o la pluma del Llobregat </dtc:text>
</dtc:sub>
</dtc:E>
<dtc:E>
<dtc:fv>
<dtc:text> pueden desviar </dtc:text>
</dtc:fv>
</dtc:E>
<dtc:E>
<dtc:cuerp>
<dtc:text> la direcciÃ³n de la corriente respecto del viento significativamente, llegando incluso a oponerse a Ã©l </dtc:text>
</dtc:cuerp>
</dtc:E>
</dtc:res_n>
</dtc:C>
</dtc:Ora>
</dtc:C>
<dtc:C>
<dtc:Ora>
<dtc:C>
<dtc:elab_s>
<dtc:E>
<dtc:conector>
<dtc:text> Sin embargo, </dtc:text>
</dtc:conector>
</dtc:E>
<dtc:E>
<dtc:cuerp>

```
                                                    <dtc:text> a escasa
carrera de marea no </dtc:text>
                                                        </dtc:cuerp>
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:fv>
                        <dtc:text> hacÃ-a esperar </dtc:text>
                                    </dtc:fv>
                                                    </dtc:E>
                                                    <dtc:E>
                                                        <dtc:cuerp>
                                                            <dtc:text> un cambio
de tendencia </dtc:text>
                                                        </dtc:cuerp>
                                                    </dtc:E>
                                                </dtc:elab_s>
                                            </dtc:C>
                                        </dtc:Ora>
                                    </dtc:C>
                                </dtc:Resultados>
                            </rdf:Description>
</rdf:RDF>
```

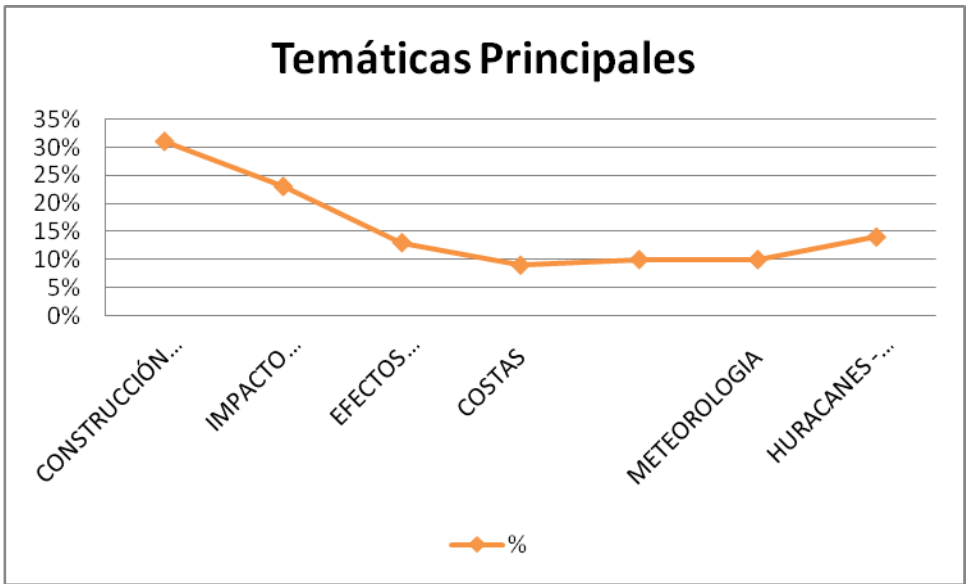
**ANEXO 37 INTEGRACION ABREVIADA DE CASOS DE USO Y ESQUEMA DE SISTEMA**

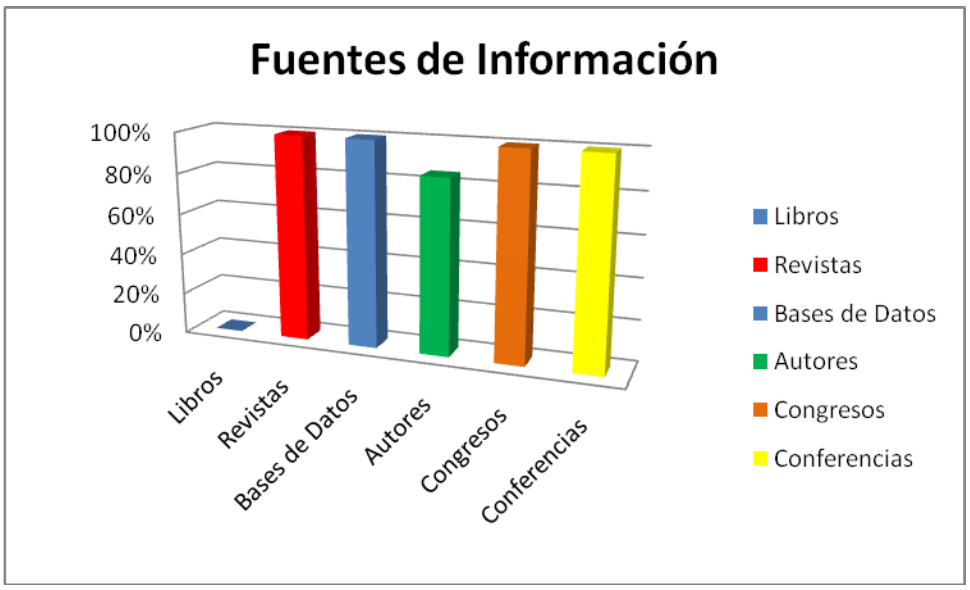


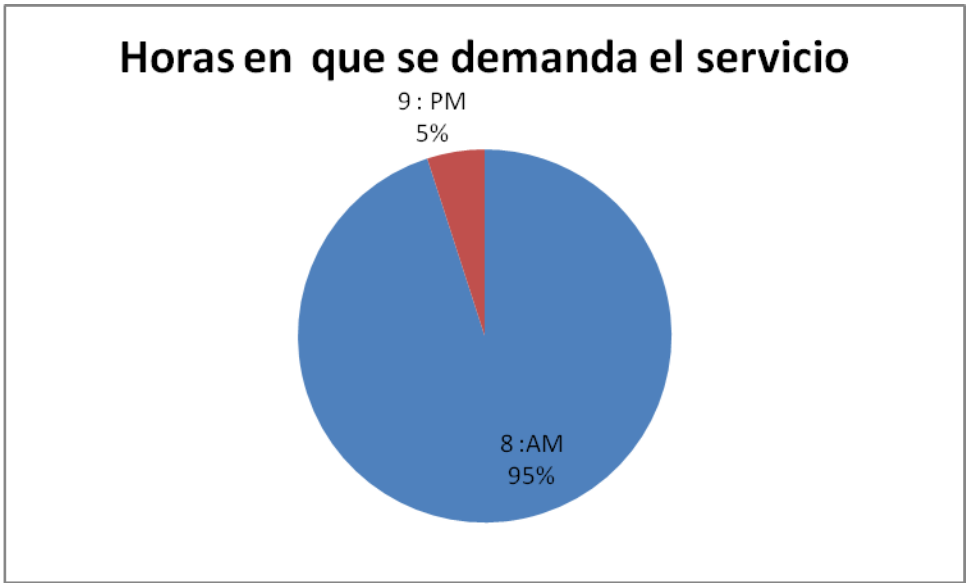
## ANEXO 38 ESTRUCTURA DEL MODELO TEXMINER

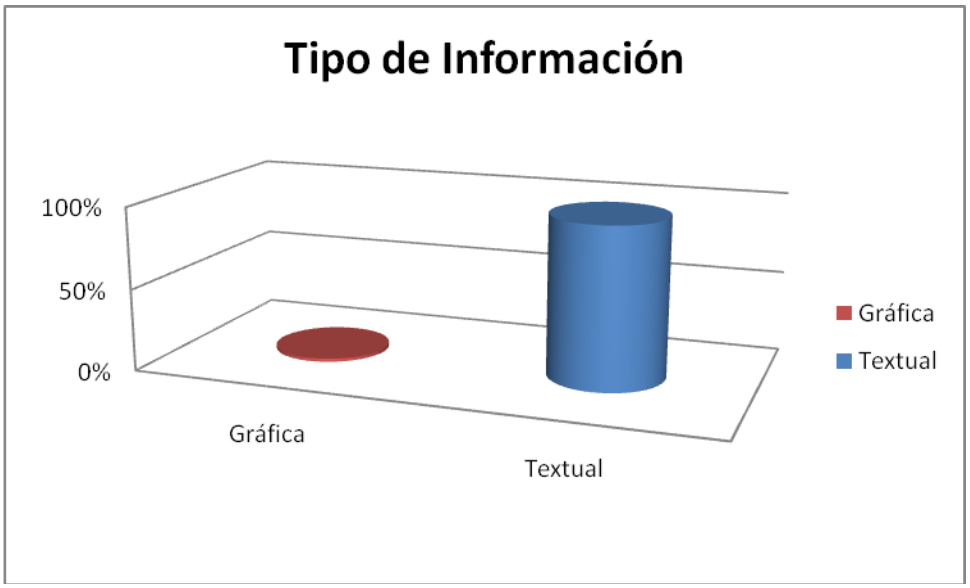


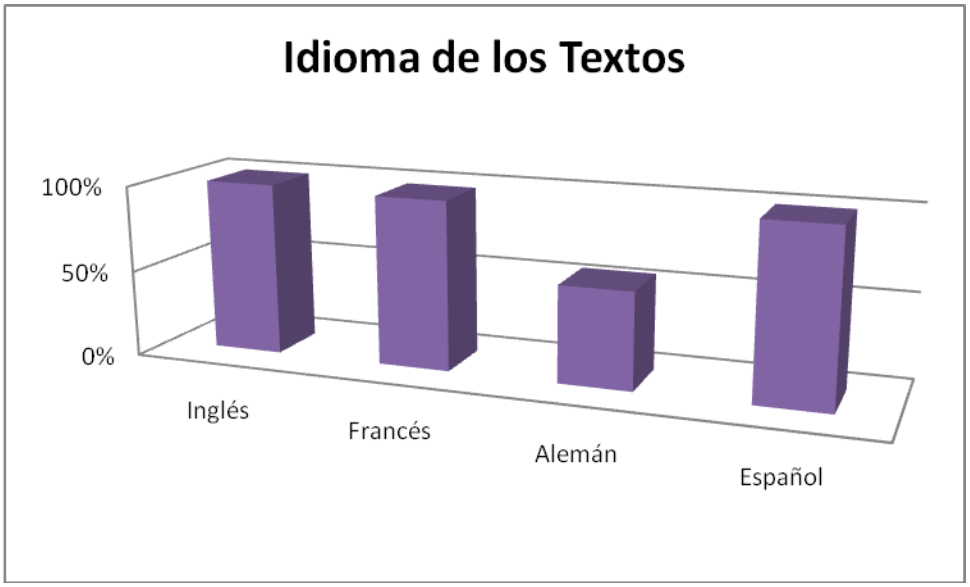
## ANEXO 39 RESULTADOS DEL ESTUDIO DE NECESIDADES




Horas en que se demanda el servicio
9 : PM
5%
8 :AM
95%


Tipo de Información
100%
50%
0%
Gráfica
Textual
Gráfica
Textual


Idioma de los Textos
100%
50%
0%
Inglés
Francés
Alemán
Español