Nuevos retos en el análisis documental de contenido: la gestión de la forma documental del contenido

Mónica Izquierdo Alonso
Universidad de Alcalá (España)

La construcción no sólo es cuestión de materiales y técnica constructiva. En ella, es preciso también el dominio del espacio, así como del conocimiento de la finalidad a la que está destinada la obra arquitectónica. El estudio de ambas permitirá la proyección (disposición y ordenación) del diseño arquitectónico, y una elección y combinación acertada de los materiales.

0.1. Resumen

Se presenta la forma documental de contenido como un desarrollo teórico novedoso dentro del Tratamiento Documental del Contenido "extendido" y se asientan las bases para la creación de un modelo formal fundamentado en los estudios textuales de género. Dentro de las implicaciones metodológicas del modelo propuesto, el concepto de estructura textual o esquema formal de contenido cobra especial importancia, a través de la identificación de las categorías retórico-funcionales y de la detección de marcadores retóricos, que poseen una función importante dentro de la teoría de la relevancia informativa. Finalmente, desde estas consideraciones pragmático-funcionales, se menciona una aplicación informática práctica que constituye una herramienta lingüística para la gestión de contenidos textuales.

Palabras clave: Pragmática documental. Análisis documental de contenido. Estructuras textuales. Lenguajes de marca.

0.2. Abstract

The documentary form of content is presented as a new theoretical development inside the "extended" Documentary Content Processing theory, for which a model is established based on the textual studies of genre. Methodologically, the concept of textual structure, or formal scheme of content, is stressed and analyzed through the identification of rhetorical-functional categories and the detection of rhetorical markers with an important function in the theory of informative rele-

vance. Finally, a practical application is presented which constitutes a linguistic tool for the management of textual contents.

Keywords: Documentary Pragmatics. Information processing. Textual structures. Markup languages.

1. Introducción

Iniciaremos el discurso centrándonos en la dimensión del contenido textual y haciendo uso del lema que encabeza este estudio. Como marco de reflexión de nuestras teorizaciones, estableceremos un símil entre el diseño arquitectónico y el discursivo. Cotejaremos, pues, la técnica de la construcción en los procesos de edificación, y las técnicas discursivas de composición textual. En este último contexto retórico-discursivo, no hablaremos de edificios materiales sino de constructos textuales. Partimos, pues, de la consideración del documento-texto como un mosaico de ideas e informaciones (materiales), en el que para su representación es preciso lograr el dominio y equilibrio del espacio textual (textualidad), de la situación comunicativa y de las técnicas de composición escrita. Se proyecta, de este modo, un documento concebido como un todo sistemático y coherente, en el que tan importante es el *qué* (el asunto, el tema), como el *modo* en que se organiza y estructura globalmente la información, en función de una situación y un propósito comunicativo concreto. Esta *estructura organizativa del contenido* es un factor esencial, que nos permite *reconocer* el tipo de información ante el que nos hallamos y valorar su relevancia textual. Dicha organización textual implica: a) la existencia de *elementos formales* constituyentes: categorías retóricas, como elementos de análisis de las estructuras esquemáticas; y b) la existencia de *relaciones funcionales* entre ellas.

Desde esta perspectiva retórico-funcional, concebimos el texto como un tejido que presenta una *estructura orgánica* a través de la cual se articula y distribuye la información. Esta disposición formal (o esquema textual característico) difiere según de qué *tipo* de documento se trate, del *propósito* de éste y de la situación comunicativa. Cabe suponer, que esta *heterogenidad estructural* del contenido informativo presenta características propias, entendidas bajo la noción de género, que han de tener su reflejo en metodologías específicas y técnicas diferenciadas de tratamiento documental de contenido (TDC). Ahora bien, desde la perspectiva "tópica" del análisis documental de contenido (ADC) no se ve representada dicha estructura, y este estudio pretende ser una intervención a favor de la consideración del *aspecto formal del contenido* del mensaje informativo, defendiendo un nuevo estatuto teórico y metodológico.

Teniendo en cuenta todo lo dicho, una vez centrada la atención en esta nueva dimensión retórica, y tras su conceptualización, nos aproximaremos a un nuevo

modelo de ADC, basado en la lingüística discursivo- funcional y en los modelos del análisis textual. En este sentido, ahondaremos en la progresión textual de un corpus lingüístico especializado, identificando los "actos" que el autor lleva a cabo para organizar y planificar su escrito —como categorías de segmentación textual—, así como los dispositivos lingüísticos que sirven de instrumento para estos propósitos. Por último, como herramienta de control y validación metodológica de los aspectos retórico-formales del contenido, aludiremos al diseño de un sistema de organización y recuperación de información, basado en ontologías y controlado por un lenguaje XML (*GeConText* v. 0.1).

2. ¿Qué ha motivado nuestro interés por los aspectos formales del contenido?

Una de las razones que han impulsado nuestra inquietud por la investigación documental de los aspectos retóricos del contenido es el necesario conocimiento que hemos de tener, como analistas de contenido, de la materia prima con la que operamos: el texto; pues preferentemente procesamos textos, información textual, aunque el ADC se abre a otros sistemas semióticos como el icónico, el auditivo, multimedia, etc. En este sentido, los textos son unidades lingüístico-comunicativas con determinadas características textuales, que han de conocerse para dominar el espacio textual y dirigir los estudios de búsqueda textual estructurada.

Consecuentemente, es necesario un dominio de las representaciones textuales y, máxime, si nuestro propósito para el tratamiento discursivo es, haciéndonos eco de la alegoría otletiana (1), aplicar "procedimientos metalúrgicos" de procesamiento documental. Con estas técnicas, ya no extraemos minerales y refinamos metales sino segmentos textuales relevantes, desde criterios depurados de estructuración textual. Ello nos lleva a la necesidad de adentrarnos en el entramando textual, y desmenuzarlo, cuidadosamente, a fin de reconocer los distintos aspectos que lo conforman. Bucearemos, así, en el conocimiento de las características textuales, los procedimientos lingüístico-discursivos y las relaciones retóricas. De aquí la necesidad de expandir cualitativamente el marco conceptual dentro del TDC, a partir de la aplicación de modelos lingüísticos, readaptados convenientemente a los intereses documentales, provenientes de campos como la lingüística textual o el análisis del discurso. Desde este marco general, intentaremos analizar y dar respuesta a dos problemas teóricos de considerable importancia y proyección en el campo del procesamiento textual automático: ¿cómo está organizada la representación del conocimiento en los documentos?; y ¿cómo se distribuyen en ellos las "partes" del conocimiento?

Otra de las motivaciones que nos dirigió la atención hacia el tema que nos ocupa fue cuestionarnos si los modelos existentes y las técnicas actuales de ADC daban respuesta a las necesidades de procesamiento informativo, creadas como

consecuencia del aumento de la información electrónica y la aparición de las redes. Dos fueron los planteamientos que nos hicimos: ¿Existen técnicas efectivas de gestión de información textual que faciliten una recuperación eficiente? ¿Se ha preocupado la Lingüística Documental tradicional de dar respuesta a este problema, desde nuevas orientaciones teórico-prácticas?

A todo ello, se suman los problemas detectados en los sistemas de recuperación de información, ante la falta de teorizaciones sobre la estructura textual y los tipos documentales. A nuestro juicio, carecemos de teorías consistentes sobre la estructura y las funciones del discurso con el suficiente grado de detalle como para realizar representaciones automáticas (generaciones automáticas de textos, búsqueda estructurada en espacios textuales diferenciados, etc.). Bien es cierto, sin embargo, que, desde campos como el procesamiento del lenguaje natural, dentro de la ingeniería lingüística, o desde la recuperación de información, se están haciendo importantes avances en el reconocimiento automático de estructuras discursivas (Marie-Francine Moens, 2000; Eduard Hovy, 1993, 1998, 2002; Daniel Marcu, 2000, 2002 etc.). En estos casos, sí se incorporan técnicas de representación lingüística, desde los actuales condicionantes del análisis discursivo. Desde este contexto, se están llevando a cabo investigaciones sobre progresiones temáticas para la elaboración de extractos y resúmenes automáticos (Paice y Jones, 1993; Lehman, 1999; etc) y, del mismo modo, han surgido una serie de estudios centrados en la semántica formal del discurso (Mann y Thompson, 1988; Maier y Hovy, 1993; Marcu,1999, 2000; etc). Las primeras se ocupan tanto del uso de áreas textuales estructuralmente significativas (párrafos de apertura, epígrafes, etc.) como de los indicadores de posición de las unidades temáticas en el discurso. En el caso de la aplicación de lenguajes de representación a la semántica discursiva se están logrando importantes avances desde la teoría de la argumentación.

Para concluir con las razones que nos llevaron hacia esta reflexión sistemática de los aspectos formales para el análisis del contenido, no podemos olvidarnos de la consideración de éstos en otras artes (como la Filosofía, la Música, la Arquitectura, la Pintura, la Lingüística, etc), con disciplinas específicas que formulan teorías y postulados sobre la relación formal (el hilemorfismo, la teoría psicológica de la Gestalt, la teoría sobre la composición musical y las formas musicales, la gramática en lingüística, y más específicamente la morfología, etc). Paradójicamente, en el tradicional análisis de contenido, no se han tenido en cuenta los condicionantes formales (estructurales) de los aspectos temáticos. Bien es cierto, que algunos autores, haciéndose eco de los avances de la Lingüística Textual, han considerado los aspectos estructurales del contenido (*vid. infra* 5.1). Sin embargo, y aunque se reconocen distintos espacios para el contenido, el ADC, tanto desde su dimensión teórica como desde su formulación metodológica, se

sigue centrando en la macroestrutura temática. No existen, pues, ni metodología ni técnicas de análisis concretas para las relaciones retórico-formales.

Esta investigación supone una llamada de atención sobre los aspectos estructurales del contenido, en las Ciencias de la Documentación, desde el reconocimiento de una nueva variante: la *Morfología Documental del Contenido* (MDC). Hemos optado por esta expresión sintagmática, como denominación que refleja el espacio de significación que pretendemos dar a los aspectos formales de contenido. Definimos, pues, la MDC como la disciplina encargada del estudio de la Forma Documental del Contenido (FDC). A ella incumbirá el establecimiento de los principios generales que rigen la estructura interna de los mensajes informativos y la delimitación de la naturaleza, relaciones y funciones de los diferentes tipos de constituyentes que pueden ser parte de dicha estructura. Dicha disciplina estaría incluida dentro de una rama más amplia: la textología documental (Izquierdo Alonso, 1999a).

3. Los ahormantes del contenido textual: género, tipo y estructura textual

Desde este espacio formal, describiremos algunos principios explicativos y operativos de la realidad que estudia, y en la que actúa, la forma o estructura documental de contenido. Distinguiremos, pues, tres aspectos como determinantes de una estructura de contenido: la adscripción a un género, la elección de una determinada forma-tipo textual, y la realización de una estructura textual concreta para cumplir determinada función comunicativa.

Un *género* es un vehículo convencional, elegido por un emisor, para transmitir todo aquello que quiere expresar. Está conformado por un conjunto de tipos, que comparten una serie de rasgos formales y de contenido, y constituye un patrón para la interacción comunicativa que se produce en un ámbito social determinado. Hablamos así de géneros musicales, géneros literarios, géneros discursivos o géneros documentales.

Descendiendo en la red jerárquica de gradación formal, nos encontramos con los términos *tipo-forma* y *estructura* textual. Ésta última constituye la textura, la urdimbre, que conforma determinados tipos de texto. Nos centraremos en la relación entre estos dos elementos de la tríada (tipo- estructura) sin cuestionarnos la conflictiva correspondencia entre géneros y tipos textuales. Para ello, haremos uso de la analogía y partiremos de una serie de ejemplos que nos ayudarán a comprender mejor no sólo la relación entre un tipo textual (forma bibliográfica/ tipo documental) y una estructura (forma/estructura documental), sino también de qué modo la función determina y condiciona la elección de una determinada estructura.

Para matizar las diferencias de significado entre los conceptos de "tipo" y "estructura" vamos a recurrir a la observación. Imaginemos varios individuos de una misma especie; los humanos, por ejemplo. Desde nuestro ejercicio mental advertiremos el aspecto o la *forma* exterior de los distintos cuerpos (*tipos*), con una serie de rasgos distintivos, como característica de cada persona. Distinguiremos así entre individuos altos-bajos, estilizados-gruesos con una tez clara-oscura; etc. La *estructura* no es visible, pero podemos conocerla a través del estudio de la anatomía. Está dentro del cuerpo, constituyendo los armazones óseos, el sistema muscular, etc. Ella va a ser la que determine la *configuración física* o la forma de un sujeto. Es decir, el que éste tenga determinadas características ligadas a la herencia genética y a las influencias del medio ambiente. En el caso de las estructuras textuales los condicionantes genéticos vendrán dados por las convenciones de género y por las exigencias del contexto situacional, como factor medioambiental. A esto, hemos de sumarle el aspecto funcional, es decir toda estructura está condicionada por la función, causa primera o motor de aquélla. Así, y siguiendo con la observación de especies biológicas, vamos a realizar un ejercicio de osteología comparada, estableciendo diferencias entre el sistema óseo de las distintas especies vertebradas (reptiles, anfibios, mamíferos, aves). Nos centraremos en la estructura morfológica de éstas últimas, y más concretamente en su aparato locomotor. Si partimos del análisis de las extremidades anteriores de un ave voladora, vemos cómo la estructura de aquéllas está condicionada por ciertas propiedades biomecánicas que responden a la necesidad del vuelo. Esta capacidad aerodinámica exigirá una cierta configuración de esta extremidad (forma, tamaño de los huesos, etc). Del mimo modo, las funciones dinámicas y prensiles de la extremidad anterior de un humano determinarán su estructura.

Esta consideración de las funciones discursivas a las que sirven determinadas estructuras sitúa nuestro estudio en una perspectiva funcional. Desde este posicionamiento, entendemos el tratamiento documental de contenido (TDC) como un tipo específico de *acción* comunicativa, insertándolo en un contexto más amplio de la teoría de la comunicación humana y de la cultura (2). Consiguientemente, concebimos el análisis de la estructura textual desde una dimensión funcional y sistémica en la que, como muestra la figura 1, tenemos en cuenta la interacción semiótica, la acción pragmática y la dimensión comunicativo textual, desde factores psicocognitivos y socioculturales. Por tanto, el modelo de FDC presentado en este trabajo, como fase constitutiva del TDC, es funcional porque atiende a una doble caracterización: a) los aspectos de la estructura discursiva, desde el punto de vista de la interacción comunicativa (*dimensión pragmática*); y b) las *funciones* que cumplen dichas *estructuras* en el desarrollo del discurso, para cumplir con los *propósitos comunicativos* de determinados tipos de textos, ante una *comunidad discursiva* dada.

4. Una aproximación al concepto de forma documental de contenido (FDC)

Podemos definir un texto como conglomerado flexible de información que admite muchas configuraciones, abarcando infinidad de materias y enfoques, según la perspectiva de su autor/obrador, la situación comunicativa en la que se inserte y la función textual a la que responda (3). Si tenemos en cuenta todos estos factores relacionados con el funcionamiento textual, concluimos que el texto no sólo transmite determinados temas, sino que también comunica con su estructura —de hecho un mismo tema puede dar lugar a distintos tipos textuales—, y ésta es algo más que un mero patrón o molde organizativo que alberga un contenido.

Todos los textos reflejan distintos niveles en su estructura y contenido, y pueden articularse en disposiciones y proporciones variadas. En consecuencia, debemos caracterizar la naturaleza de esas formas elementales o básicas de organización de la información, y profundizar en la definición y características más comunes a los diferentes tipos textuales, así como en las relaciones y funciones que integran dichas estructuras. Esto es tarea de la disciplina referida anteriomente como Morfología Documental del Contenido (MDC).

La FDC (4), como objeto de estudio o materia prima sobre la que trabajaría dicha disciplina, constituye el armazón interno, que vertebra el texto, el entrama-

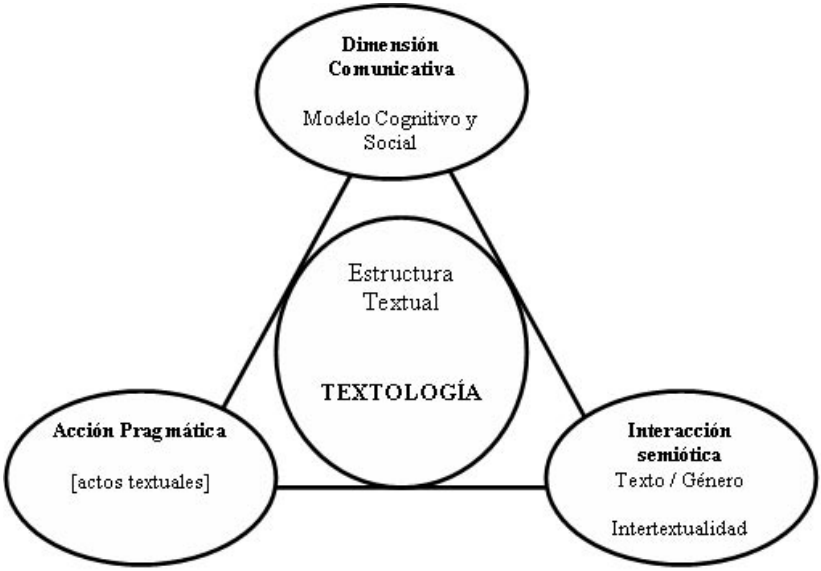

Fig. 1: Las dimensiones de la estructura textual.

do lógico-retórico revestido de conceptos, de contenidos temáticos, a los que da sustrato. Este entramado es lo que caracteriza los tipos de texto (independientemente del contenido semántico) y diferencia unos géneros textuales de otros. Así las cosas, el mensaje o contenido puede ser considerado bajo dos perspectivas complementarias: como forma y como sustancia. Consiguientemente, en el TDC no hemos de referirnos sólo a la naturaleza del contenido semántico (tema). Es preciso atender también a la red lógico-retórica de relaciones que configuran, sustentan y conformar las estructuras temáticas. Es dicho armazón lo que hace posible la materialización de la estructura semántica, la perceptibilidad integral y la coherencia global del texto.

5. Metodología del diseño experimental: el modelo formal de análisis

Fundamentado el contexto teórico en torno a los aspectos formales o retóricos del contenido, desarrollaremos un marco analítico que toma como referencia la perspectiva sistémico-funcional de análisis textual. Dicho modelo se aplicará al análisis de estructuras (formas) retóricas de un tipo documental específico: el artículo científico. Así, a partir de un análisis de corpus, se diseñara una gramática textual basada en de la detección de una serie de constituyentes (categorías morfofuncionales) y en la proyección de reglas formales. Finalmente, este modelo formal encontrará su representación en una ontología, en la que se representará jerarquizada la estructura retórica del documento-tipo analizado.

5.1. Precedentes de nuestro modelo: enfoques textuales y recuperación de información

Sin ánimo de ser exhaustivos, vamos a mostrar una breve relación de distintas propuestas sobre representación del conocimiento y recuperación de información, desde los avances conseguidos en las teorías lingüísticas y en el dominio de la Ingeniería informática. Así pues, podemos aproximarnos al estudio del procesamiento informativo a partir de distintos objetivos:

- El funcionamiento de la coherencia textual en el marco de investigaciones sobre traducción automática, planificación textual y elaboración de resúmenes (Marcu, 1998, 2000; Hovy, 1993, 1995, etc.).
- La búsqueda de parámetros que sustenten juicios evaluativos sobre los procesos de calidad en el resumen y en la indización: revisión de las condiciones de textualidad (Pinto, 1994; Moreiro, 2002)).
- Las teorías sobre el procesamiento lector, adaptadas a los fenómenos de la indización: tipologías textuales, metodología de estratificación textual en micro, macro y superestructuras (Pinto, 1996, 2001; Moreiro, 1993, etc.).

- La perspectiva informática centrada en las estructura lógicas, modelado de datos y aplicación de estándares SGML/XML (Nogales et al., 2003, etc.).
- La elaboración de un *modelo de estructuración de contenidos*, desde la perspectiva sistémico-funcional del análisis de género. Representación de espacios formales de contenido y búsqueda estructurada de contenidos. (Izquierdo Alonso, 2002, 2003).

Desde este último enfoque, considerando la heterogeneidad teleológica y la consiguiente pluralidad de modelos lingüísticos anteriores, justificaremos nuestro modelo pragmático de representación documental del contenido.

5.2. Diseño de un modelo de estructuración de contenidos para la representación documental.

El TDC es una operación de metacomunicación (5), una práctica semiótica con una manifestación textual importante. De ahí, que tanto su conceptualización como su ejercitación operativa hayan de tener un enfoque textual, dentro de sus posibles realizaciones semióticas. Desde esta consideración, nos aproximaremos a otras disciplinas para las que el texto, y su entorno, es objeto de reflexión. En esta línea argumental, dirigiremos nuestra mirada hacia la lingüística textual (LT) y al análisis del discurso (AD), tanto desde sus áreas de investigación teórico-descriptivas como desde su práctica empírica. En el primer caso, la investigación textual se halla más centrada en los factores internos al texto (condiciones de la textualidad, tipologías, gramáticas textuales, etc). En el segundo, desde el análisis discursivo, la actuación se centra en una perspectiva comunicativa o contextual, incorporando al estudio del texto las condiciones de producción y recepción, entre otros condicionantes. Así considerado, desde esta perspectiva textual, el TDC quedará caracterizado como un *acto* complejo *de comunicación* (se superponen tres tipos de discurso), en el que intervienen no sólo factores internos al propio texto, sino modelos de comunicación y condicionantes socioculturales de índole extratextual.

Desde la revisión de la metodología textual, encontramos diversos modelos descriptivos del texto. Esta heterogenidad conlleva desarrollos teóricos y metodologías de análisis diferentes: estudios sobre tipologías textuales (Beaugrande y Dressler, 1981; Werlich, 1976; Ciapuscio, 1994; Combettes, 1988; Charolles, 1988; etc.); modelos semánticos macroestructurales (Van Dijk, 1983, 1990, 1996); análisis de relaciones lógico-retóricas (Meyer, 1984); definición de secuencias prototípicas (Adams, 1992)); formulación de arquetipos discursivos (Bronckart, 1996)); enunciación de esquemas secuenciales (Roulet, 1989); etc.

Serán los modelos sencuenciales de Roulet y los arquetipos de Bronckart los que constituirán la base de nuestro modelo pragmático, aplicado al TDC. Este *enfoque funcional* se caracteriza por el papel que ejerce la función textual dentro

del análisis del texto origen, así como la función textual de la representación documental en el texto- producto, resultante del proceso documental.

5.2.1. Actos de habla y función textual

En nuestro modelo de análisis formal concebimos el texto como un "artefacto" planificado con una orientación pragmática. Consiguientemente, la unidad de segmentación textual será el *acto discursivo*. Un texto podrá definirse, pues, como una sucesión jerárquica de actos discursivos. Aplicamos el término *función* para designar cualquier *actividad humana* realizada lingüísticamente (exhortar, preguntar, explicar, informar, describir, anunciar, etc.), en coincidencia con la teoría de los actos de habla de Austin (1962) y Searle (1969). La importancia otorgada a la *función textual* constituirá el fundamento de la *corriente funcionalista* que defendemos para el TDC. Éste se da dentro de un marco social y sus productos cumplen una determinada función, según el contexto en el que se encuentren inmersos. Es esta dimensión funcional la que diferencia, fundamentalmente, los distintos productos de este tratamiento documental. Pero esta noción de función no atiende sólo al producto resultante, se da igualmente en todo el proceso de tratamiento, y afecta también a la metodología general del análisis. Ello nos lleva a ligar la función con la definición de los distintos tipos textuales, y hacia una metodología específica para cada uno de los géneros documentales.

El concepto de *función* es, pues, un elemento decisivo para nuestro modelo de análisis, ya que va a orientar y condicionar la transformación de contenidos que tiene lugar el la representación documental, dependiendo del tipo de actantes que intervienen en la interacción documental.

5.2.2. Cuestiones de arqueología textual: la metodología funcional del análisis textual

Para ilustrar nuestro enfoque sistémico de estructuración de contenidos, referido en el epígrafe anterior, partiremos de una comparación entre la técnica funcional de análisis textual y la metodología arqueológica. Así, al igual que la arqueología estudia los restos materiales, producto de la cultura humana, nuestra perspectiva de análisis formal, desde la *explotación* y estratificación del enclave textual, intenta explorar las huellas de los actos comunicativos. En este caso, tal y como muestra la figura 2, los restos son estructuras retórico-textuales. Tomamos éstas como elementos activos de interacción y, al mismo tiempo, como vestigios de las distintas etapas de la planificación textual que nos ayudan a reconstruir el texto original, desde sus intenciones u objetivos retóricos. El concepto de *acto discursivo*, jerarquizado en secuencias y movimientos, como jalones de estructuración textual, se convierte en algo central, en un intento de identificar en la progresión textual de un corpus, los "actos" que el autor lleva a cabo para organizar

y planificar su escrito y los dispositivos lingüísticos que sirven de instrumento para estos propósitos.

5.3. Presupuestos básicos del análisis

Desde la correlación significativa entre forma y función textual, proponemos un modelo de análisis para el TDC que combina dos tradiciones de investigación. Nos referimos a la teoría glosemática danesa de estructuración del signo lingüístico y al análisis de género, como marco teórico-metodológico en el que se describen las categorías del modelo pragmático-textual propuesto.

5.3.1. La teoría glosemática danesa y la concepción Hjelmsleviana del signo lingüístico.

Antes de abordar el análisis del aspecto formal del contenido haremos una breve presentación del esquema y las aportaciones teóricas sobre las que hemos

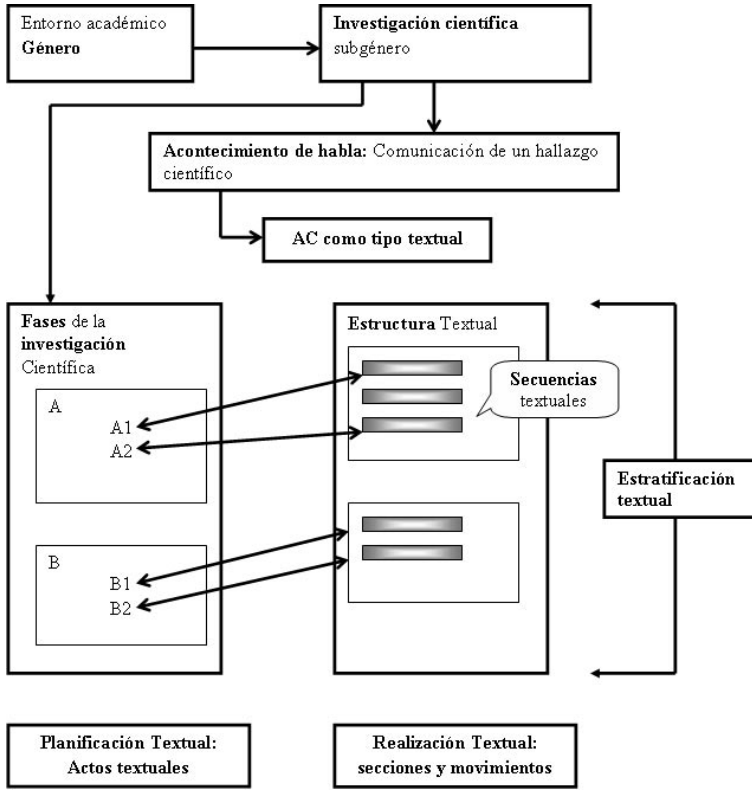

Figura 2: Perspectiva funcional del análisis de género de un artículo científico (AC).

basado los principios de nuestra hipótesis: el esquema hjemsleviano para el estudio del signo lingüístico (Izquierdo Alonso, 1999b, 2000; Izquierdo Arroyo, 1993). Uno de los postulados característicos de la escuela danesa, o glosemática, es la consideración de que el contenido, al igual que su expresión, debe ser contemplado bajo dos aspectos complementarios: como forma (estructura) y como sustancia. Frente a la distinción dicotómica saussureana del signo (Expresión= significante; contenido= significado), Hjelmsliev propone una caracterización del signo basada en un modelo que atiende a la superposición tetrapartita de planos y niveles. Como se puede observar en la figura 3, distingue dos planos: el de la expresión y el del contenido, a cada uno de los cuales se atribuye correlativamente una sustancia y una forma.

Contenido	Sustancia
	Forma
Expresión	Forma
	Sustancia

Figura 3. Planos y niveles en el signo lingüístico.

Identificaremos, pues, una sustancia y una forma de la expresión, y una sustancia y una forma del contenido. En el *estrato semántico-cognitivo*, la *sustancia del contenido* textual (SC) está constituida por todo el conjunto de ideas, hechos, pensamientos, etc. pertenecientes al continuum del espectro semántico-cognitivo (el universo ideológico). Mediante la *Forma del contenido* (FC), y según la intención del autor, esa sustancia se materializa en una estructura lógico-retórica definida, en unos modelos de organización textual, y el texto adquiere un orden y disposición característicos. A su vez, en el *estrato gráfico lingüístico*, en el *plano de la expresión*, el conjunto heterogéneo de grafías de un texto (letras, esquemas, ilustraciones, notaciones, gráficos) que constituyen la sustancia significante cristalizan en una estructura formal precisa, que adquiere una determinada configuración o disposición gráfica en torno a una página o cualquier otra superficie suscriptora. Hecha esta distinción, nos situamos en el estudio del *plano del contenido*, núcleo y objeto propio del TDC, distinguiendo en él dos unidades interdependientes de análisis: a) la sustancia del contenido, y b) la forma del contenido.

5.3.2. El análisis sistémico de género

El otro pilar sobre el que se asienta nuestro modelo de arqueología textual es el concepto de género (8). Autores como Bathia (1993), Swales (1981,1990), o Paltridge (1997), consideran los géneros como tipos de textos (orales y escritos)

definidos por sus propiedades o características formales, así como por sus propósitos o intenciones comunicativas, dentro de un contexto social. Esta perspectiva de género incluye una dimensión pragmática y sociocultural en las investigaciones sobre caracterización estructural de los textos y ha asumido el modelo de análisis contextual de la escuela de Sydney. Lo más destacable de la teoría del análisis de género es que ha asentado las bases para el análisis estructural desde el punto de vista pragmático, delimitando una estructuración retórica de los textos en fases (*moves)* y pasos (*steps*) atendiendo a las funciones y subfunciones comunicativas concretas de género. Del mimo modo, supone un modelo sociopragmático de análisis textual basado en la interacción textual (Izquierdo Alonso, 2003a).

Nuestra investigación es eminentemente pragmático-discursiva. Como mostraremos en el siguiente epígrafe, partimos de un corpus textual especializado y analizamos sus características propias y su configuración intratextual, con el objetivo de obtener su estructura. Para ello, atendemos preferentemente a la relación que existe entre los propósitos comunicativos e interactivos de cada secuencia textual y a los medios que se ponen en juego para realizarlos.

5.4. Propuesta de análisis sistémico de un género específico: el artículo científico (AC)

Una vez definida nuestra orientación metodológico-funcional, y delimitados los objetivos del modelo pragmático de análisis textual (Izquierdo Alonso, 2003b, p. 749-750), intentamos aplicarlo a un género concreto, el científico-académico. Nos centramos en un tipo específico, el artículo científico, y analizamos su organización retórico-discursiva a partir de un corpus de veinte textos (6), pertenecientes al ámbito de la Psicología experimental. En nuestro análisis, partimos del modelo esquemático de Swales (1990) para la división retórica del artículo científico. El autor concibe una estructura argumentativa, fija y obligatoria, fundamentada en cuatro grandes secciones (MIRAD) que contribuyen al desarrollo lógico del artículo. Asimismo, existen otra serie de elementos opcionales, como constituyentes inmediatos de cada sección: los *moves* y *submoves*. Dichas variables desempeñan una serie de funciones discursivas específicas que completan el valor informativo de cada sección textual y contribuyen a la realización global del objetivo comunicativo de este tipo textual: la comunicación-informe de una actividad investigadora.

Así pues, en nuestro esquema de organización discursiva, distinguimos dos tipos de unidades (actos textuales) y una serie de dispositivos retórico-lingüísticos que permiten la realización y el reconocimiento posterior de aquellos (Izquierdo Alonso, 2002): una serie de macroactos textuales o macrocategorías: (secciones textuales o zonas primarias; los movimientos y submovimientos retóricos (categorías retóricas o morfofuncionales) como microactos o secuencias textuales; y

los marcadores discursivos como dispositivos lingüísticos o "señalizadores" de las transacciones comunicativas.

Para la identificación general de los movimientos estructurales, realizamos un análisis retórico-comparativo a partir de las clasificaciones de 16 autores. El estudio nos reveló que las secciones más estudiadas, y en las que con mayor claridad se mostraban los *moves* o las estrategias retóricas (dependiendo de la terminología usada por cada autor) eran las de "introducción" y "conclusiones". Para completar esa visión inicial (y, al mismo tiempo, cubrir la estructura poco estudiada de secciones como la "metodología" o "resultados"), recurrimos a los trabajos existentes sobre metodología general de la investigación científica. En ellos se aludía a fases y subfases concretas de la actividad investigadora que bien podían relacionarse con los movimientos retóricos como informe de dichas actividades.

5.4.1. Categorías retórico funcionales

A partir de las definiciones estructurales y de la identificación de *moves*/*submoves* con sus correspondientes funciones, delimitamos 547 categorías retóricas o espacios formales de contenido. Estos esquemas se manifiestan como parte de la estructura formal del texto, conformando una red jerárquica de *actos discursivos* cuyos elementos constituyentes son las secciones textuales y los encadenamientos discursivos (*moves-submoves*) como especificaciones funcionales de aquéllas. Desde la *perspectiva sistémico-funcional* responden a la necesidad de caracterizar la información en segmentos textuales atendiendo a la naturaleza y dimensión comunicativa del texto. Cada categoría proporciona al lector distintos tipos y cantidades de información. Desde la *dimensión documental* constituyen zonas de información a partir de las cuales podemos organizar y recuperar el contenido de los textos pertenecientes a un género documental y tipo textual concreto.

Para la detección apriorística de las categorías en nuestro corpus nos servimos de marcadores discursivos u operadores pragmáticos. Estos dispositivos textuales actuaban como balizas que orientaban nuestra búsqueda de las funciones discursivas dentro de cada *movimiento* de la estructura textual. Es decir, se erigían en señalizadores de "cambio de *move*" o "inicio de *submove*", como identificadores de categorías retórico-funcionales en nuestra muestra textual.

Una vez que teníamos identificados los constituyentes, necesitábamos un gramática textual que nos definiese las relaciones entre ellos. Apoyándonos en los fundamentos de la gramática chomskiana generativo- transformacional, y mediante un sistema de reglas de reescritura, obtuvimos la jerarquización categorial y un total de 171 categorías terminales a las que debían asignarse marcadores. Dichas reglas constituyeron la base para la generación de una Ontología (ONTOFORM) como uno de los modelos de representación de la Forma Documental de Contenido. En dicha ontología los nodos estaban constituidos

por cada una de las categorías morfofuncionales. Además, el módulo Ontoform nos permite validar y retroalimentar el sistema de selección de categorías para un tipo textual y la asociación de marcadores pragmáticos para cada categoría.

5.4.2. Marcadores retórico-discursivos u operadores pragmáticos de función documental

Ponemos énfasis en la condición *pragmática* (7) del marcador discursivo, cuya misión es organizar y estructurar un determinado tipo de encadenamiento o *secuenciación discursiva* entre enunciados; erigiéndose así en reclamo de la relevancia del contenido informativo. Los marcadores extraídos de la pequeña muestra textual, y religados a determinadas categorías retóricas, fueron muchos (1307, 1193 depurados).

Respecto a la metodología de selección, asignación y validación de marcadores, hubo una primera fase de *extracción manual* de *frases significativas* que sirvieron para identificar las categorías retórico-funcionales detectadas a priori: método de *selección por asignación* a una categoría retórico-funcional. A partir de esta selección contamos con una muestra representativa de marcadores para ser asignados a las categorías. En una segunda fase, se produjo la asignación y validación de marcadores (candidatos y/o efectivos) a categorías terminales de un modo controlado por un procedimiento automático (componente *Remarcat* del módulo de Gestión Ontologías formales (*ONTOFORM*)**.**

Tras el análisis de estos marcadores o fórmulas, como expresiones compuestas tomadas literalmente del corpus textual, establecimos una distinción entre los macadores continuos —aquellos cuyos componentes (palabras) no tienen ningún vacío o variable: "en la línea de", y que podemos clasificar en continuos *literales* (puros o lematizados) o *con variables* en cuanto a la forma ordinal— y marcadores discontinuos —aquellos en los que localizamos vacíos o variables: "*por ... entendemos*", etc. Actualmente, nos hallamos en la fase de desarrollo expansivo del tratamiento de los marcadores; que sin duda requeriría un estudio más profundo dentro del ámbito de la Lingüística computacional. Dependiendo del nivel de detalle de análisis en los marcadores, podremos lograr una *red de sustitución funcional* y *sintáctica* que pueda incorporarse a un sistema experto de conocimiento.

6. Un escenario de aplicación: GECONTEXT como entorno de gestión de la forma documental del contenido.

Finalmente, no queremos cerrar este espacio sin aludir, aunque sea objeto ya de otro *discurso*, al modelo de aplicación que justifica y corrobora la validez general del modelo teórico-metodológico aquí presentado. Así pues, hemos desarrollado una serie de instrumentos de análisis automático creados *ad hoc*,

que permiten explorar este dominio de la FDC de un modo más representativo y eficaz. En este contexto, se ha diseñado e implementado un entorno informático de gestión de contenidos textuales (*GeConTex*t) que permite generar estructuras de conocimiento eficaces, a partir de información no estructurada (texto llano o crudo), y es capaz de categorizarlas desde parámetros no sólo basados en análisis estadísticos de RI sino en criterios discursivos de análisis textuales previos. El sistema nos posibilita el reconocimiento automático de la estructura de un determinado tipo documental, su modelización a través de una ontología y la navegación por las zonas formales de contenido en un entorno XML. Asimismo, nos permite realizar búsquedas específicas de información en estos espacios formales de contenido, complementando, los actuales sistemas de extracción y recuperación de información (incluida la de los llamados "documentos estructurados"). Todo ello con miras a ulteriores desarrollos en el campo del tratamiento documental de contenido y la recuperación de información. Esto asienta un pilar básico y fundamental para la creación de unos sistemas más eficaces de recuperación de información, al identificar zonas formales de contenido textual, desde los parámetros del análisis del cotexto y contexto lingüístico; hasta el momento inexploradas, tanto en el área documental como en el mundo de la inteligencia artificial.

7. Conclusiones

Para concluir este trabajo, queremos hacer una serie de reflexiones generales que intentan aunar los principios teóricos, base epistemológica de nuestro modelo, con la propuesta metodológica de análisis funcional, como marco operativo de aquélla. Así pues, defendemos que este nuevo enfoque formal, desde la óptica de la glosemática danesa y las actuales tendencias de análisis del discurso, enriquece y abre una nueva vía dentro de la Lingüística Documental capaz de explicar muchos fenómenos referidos al contenido. Supone ampliar el campo del tratamiento documental de contenido a la consideración de la *forma documental del contenido* y adecuar a ello los instrumentos y las técnicas de análisis a ello. Se plantea, así, la necesidad de establecer una tipología documental atendiendo al "contenido", potenciando el estudio de los distintos "géneros", en razón de su estructura discursiva. Este espacio descriptivo y teórico tiene una aplicación práctica concreta al trasladarse a un análisis de corpus de textos de especialidad, aplicando los principios de la teoría estándar de las gramáticas generativas, gracias a la definición de una base de categorías morfofuncionales y una serie de reglas de proyección. Dicha gramática de relaciones retóricas encuentra su espacio de representación en la implementación de una ontología formal para un género científico específico: el artículo de investigación.

Considerado desde esta perspectiva funcional, el análisis de género supone un avance más dentro de la investigación sobre la Forma Documental del

Contenido (FDC), una orientación entre las posibles perspectivas de análisis que se abren, creemos, dentro de este prometedor campo de la morfología documental del contenido (MDC).

8. Notas

(1) Otlet (1934a, 373-373 bis, nº 411.2), citado por el profesor Izquierdo Arroyo (1995: 35) refiriéndose al concepto y propósitos de la Documentación, hace la siguiente reflexión sobre el proceso de la *metalurgia documental*: "Es preciso poner orden en la montañas de papeles y documentos: es preciso crear una metalurgia del papel, hacer galerías de aproximación hacia esas montañas, cuyos flancos encierran tesoros; extraer de ahí el buen mineral y separar después el metal puro de su ganga (…)".

(2) Para la *concepción pragmática del discurso documental*: *vid*. el *modelo semiótico documental* del prof. Izquierdo Arroyo (1993). Esta dimensión pragmática del TDC, desde el condicionante retórico-comunicativo, se completa con la visión de Izquierdo Alonso (1999b, 2000).

(3) Así, dentro de un dominio científico las funciones pueden ser varias: difundir ampliamente un nuevo hallazgo ante la comunidad científica, transmitir el conjunto de saberes propios de una disciplina, divulgar unos conocimientos básicos ante un público de carácter general, etc. Por su parte, en una situación o contexto técnico, por ejemplo, dentro del dominio industrial, las funciones a las que ha de responder un texto pueden ir desde asegurar los derechos jurídicos frente a posibles plagios, caso de las patentes, hasta incitar a la compra de un producto si se trata de textos publicitarios.

(4) El término "forma", y el concepto que éste entraña, es un tema de estudio para diferentes disciplinas que consideran el espacio semiótico-textual como marco de actuación. Cada una de éstas lo ha adaptado a sus necesidades y, lo que es más importante, lo ha dotado de connotaciones específicas, sin quizás molestarse en un análisis interdisciplinar detallado del término y de su significación. Así, áreas como la Psicología cognitiva, la Inteligencia artificial, aprovechando los desarrollos de aquélla, la Lingüística textual o la Informática, entre otras, han definido el concepto de "forma" bajo los términos de "superestructura", "superestructura esquemática", "esquemas de dominio", "estructura retórica", "marco" (*frame*), "etiqueta", "organización lógico-discursiva", etc. El resultado lógico es un gran variedad de términos, que bien pueden no significar lo mismo, y que, sin embargo, dan buena cuenta de la confusión terminológica al respecto. Ante la señalada falta de comparabilidad de conceptos teóricos y operacionales, proponemos una acepción del concepto de forma, desde la aplicación documental de los presupuestos hjelmslevianos, que puede constituir un primer paso para un trabajo interdisciplinar más integrado y acumulativo en este dominio de los espacios formales de representación estructural.

(5) Sobre el macroprocreso del TDC como acto comunicativo de reconducción del discurso *Vid*. Izquierdo Arroyo (1993, p. 206).

(6) Tomamos el concepto de "análisis de corpus" en el sentido metodológico del valor que tiene el conjunto de los textos codificados como materia prima y banco de pruebas para la confirmación de nuestra hipótesis de estructuración textual. Del mismo modo,

el análisis de corpus constituye una herramienta lingüística de primer orden para la creación de una gramática formal que nos permita diseñar un sistema de gestión de contenidos textuales.

(7) Los marcadores no sólo pueden darse a *nivel local* (cohesivos) sino que también pueden formar parte de la macro-organización discursiva. En este caso, este tipo de operadores, de naturaleza pragmática puesto que se dan a *nivel de actos de habla*, actúan como marcadores de las categorías retórico-funcionales. Junto a este tipo de marcadores formales existe también otro tipo de marcadores temáticos: (introducción de tema, mantenimiento, reparación, recuperación de tema, etc).

9. Referencias

Anscombre, J. C.; Ducrot, O. (1988/1994). La argumentación en la lengua. Madrid: Gredos, 1994.

Austin, J. L. (1962). How to do things with words. // J. O. Urmson (ed). Londres, Oxford University Press, 1962.

Beaugrande, R. A. de ; Dressler, W. U. (1997). Introducción a la lingüística del texto. Barcelona: Ariel, 1997.

Bhatia, V. K. (1993). Analysing Genre: study of its application to professional genres. London: Cambridge University Press, 1993.

Bronckart, J. P. (1985). Le fonctionnement des discours. Um modèle psychologique et une méthode d'analyse. Neuchâtel: Delachaux et Niestlé, 1985.

Charolles, M. (1988). Les plans d'organization textuelle: périodes, chaînes, portées et séquences.// Practiques. 57 (1988) 3-13.

Ciapuscio, G. E. (1994). Tipos textuales. Buenos Aires: Universidad, 1994.

Combettes, B. (1988). Pour une grammaire textuelle. La progression thématique. Bruselas/París: De Boeck/Duculot, 1988.

Hovy, E. H. (1993). Automated Discourse Generation Using Discourse Structure Relations // Artificial Intelligence. 63:1-2. Special Issue on Natural Language Processing. 1993. 341-386.

Hovy, E. H. (1995). The multifunctionality of discourse markers. // Workshop on discourse markers. Egmond-aan-Zee, The Netherlands, 1995.

Hovy, E.H. and C.Y. Lin. (1998). Automated Text Summarization in SUMMARIST. // M. Maybury and I. Mani (eds). Advances in Automatic Text Summarization. Cambridge: MIT Press. 81-110.

Hovy, E.H. (2002). Automated Text Summarization. // R. Mitkov (ed). Oxford University Handbook of Computational Linguistics. Oxford: University Press, 2002.

Izquierdo Alonso, M. (1999a). Una aproximación interdisciplinar la estudio del usuario de información: bases conceptuales y epistemológicas. // Investigación bibliotecológica. 13: 26 (1999) 112-134.

Izquierdo Alonso, M. (1999b). Forma del contenido y función documental: el papel de la estructura en la organización y representación del conocimiento. // Actas de IV Congreso ISKO-España: representación y organización del conocimiento en sus

distintas perspectivas: su influencia en la recuperación de información. Granada: Universidad, 1999. 47-52.

Izquierdo Alonso, M. (2000). Nuevos enfoques en el estudio del TDC desde los presupuestos de las Ciencias del Lenguaje. // Scire. 6:1 (2000) 143-163.

Izquierdo Alonso, M. (2002). La forma documental de contenido: un modelo para su representación. Granada: Universidad de Granada, Dpto. de Biblioteconomía y Documentación, 2002. Tesis doctoral.

Izquierdo Alonso, M. (2003 a). Procesamiento pragmático para el tratamiento documental de contenido. // Documentación de las Ciencias de la Información. 26 (2003). 18 p. (En prensa).

Izquierdo Alonso, M. (2003 b). El análisis de género como metodología para la organización y representación del conocimiento. // José Antonio Frías y Críspulo Travieso (coords). Tendencias de investigación en Organización del Conocimiento. Salamanca: Universidad, 2003. 747-754.

Izquierdo Arroyo, J. M. (1993). De la semiótica del discurso a la semiótica documental. // Moreiro González, J.A.: Aplicación de las Ciencias del texto al resumen documental. Madrid: Universidad Carlos III; Boletín Oficial del Estado, 1993. 199-216.

Izquierdo Arroyo, J. M. (1995). La organización documental del conocimiento. Madrid: Tecnidoc, 1995.

Lehmam, A. (1999). Text Structuration Leading to an Automatic Summary System: RAFI. // Information Processing and Management. 35:2 (1999) 181-191.

Maier, E. and Hovy, E. (1993). Organizing Discourse Structure Relations using Metafunctions. // H. Horacek and M. Zock (eds). New Concepts in Natural Language Generation: Planning, Realization, and Systems. London: Pinter. 69-86.

Mann, W. C.; Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. // Text. 8:3 (1988) 243-281.

Marcu, D. (1999). Discourse structure, rhetorical parsing and text summarization [En línea]. URL: <http://www.cs.toronto.edu/compling/Topics/Discourse.html> . Consultado el 17/6/2003.

Marcu, D. (2000). Extending a Formal Computational Model of Rhetorical Structure Theory with International Structures. // The 18th International Conference on Computational Linguistics. COLING'2000. Saarbrueken, 2000.

Marcu, D.; Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Conference. Philadelphia, PA, 2002 (July). 7-12.

Marcu, D. (1998). The rethorical parsing, summarization, and generation of natural language texts. Toronto: University of Toronto, Department of Computer Science, 1998. PhD thesis.

Martín Galan, B.; Rodríguez Mazo, D. (2000). Estructuración de la información mediante XML: un nuevo reto para la gestión documental. // La gestión del conocimiento: retos y soluciones de los profesionales de la información. Actas de las VII Jornadas

Españolas de Documentación. Bilbao: Fesabid; Universidad de Deusto, 2000. 113-123.

Meyer, B. J. (1984). Text dimensions and cognitive processing. // Mand, Stein, y Trabasso (eds). Learning and comprehensión of texts. Hillsdale, NJ: Erlbaum, 1984.

Moens M. F.; Uyttendaele, C.; Dumortier, J. (2000). Intelligent Information Extraction from Legal Texts. // Information & Communications Technology Law. 9:1 (March 2000) 17-26.

Moreiro González. J. A. (1993). Aplicación de las Ciencias del Texto al Resumen Documental. Madrid: Universidad Carlos III; BOE, 1993.

Moreiro González. J. A. (2002).: Criterios e indicadores para evaluar la calidad del análisis documental de contenido. // Ciencias de la Información. 31:1 (2002) 53-60.

Nogales Flores, J. T.; Martin Galán, B.; Arellano Pardo, M. C. (2003a). Una propuesta para el tratamiento documental de las resoluciones judiciales en España haciendo uso de tecnologías XML. // Los sistemas de información en las organizaciones: eficacia y transparencia. Actas de las VIII Jornadas Españolas de Documentación. Barcelona: Fesabid, 2003. 385-393.

Nogales Flores, J. T.; Martin Galán, B.; Caridad Sebastián, M. (2003b). Una experiencia de aplicación de XML y TEI a obras teatrales del Siglo de Oro. // Los sistemas de información en las organizaciones: eficacia y transparencia. Actas de las VIII Jornadas Españolas de Documentación. Barcelona: Fesabid, 2003. 395-402.

Paice, C. D.; Jones, P.A. (1993). The identification of important concepts in Hhghly structured technical papers. // Proceedings of the 16 th ACM/SIGIR. (1993) 69-78.

Paltridge, B. (1997). Genre, frames and writing in research settings. Amsterdam; Philadelphia: John Benjamins, 1997.

Perelman, Ch.; Olbrechts-tyteca, l. (1989). Tratado de la argumentación: La nueva retórica. Madrid: Gredos, 1989.

Pinto, M. (1994). Indicadores de calidad descriptiva en la gestión de los procesos analíticos-documentales. // Actas de las Jornadas españolas de documentación. Gijón, Oviedo: Universidad, 1994. 184-204.

Pinto, M.; Gálvez, C. (1996). Análisis documental de contenido. Madrid: Síntesis, 1996.

Pinto, M. (2001). El resumen documental: paradigmas, modelos y métodos, 2ª ed. Madrid: FGSR, 2001.

Roulet, E. (1989). Modèles du discours. Berna: Peter Lang, 1989.

Searle, J. R.: Speech Acts. Londres, Cambridge University Press, 1969. Trad. cast. de Luis M. Valdés Villanueva: Actos de habla. Madrid: Cátedra, 1980.

Swales, J. M. (1990). Genre analysis English in academic and research settings. Cambridge, University Press, 1990.

Van Dijk, T. A. (1983). La ciencia del texto. Un enfoque interdisciplinario. Barcelona: Paidos, 1983.

Van Dijk, T. A. (1990). La noticia como discurso. Barcelona: Paidos, 1990.

Van Dijk, T. A. (1996). Discourse as structure and process. London: SAGE, 1996.

Werlich, E. (1976). A text grammar of English. Heidelberg: Quelle and Meyer, 1976.