EL LENGUAJE DE INTERROGACIÓN: UNA GRAMÁTICA FORMAL PARA LA RECUPERACIÓN DE INFORMACIÓN

Mario Pérez Gutiérrez*

Resumen: El conocimiento del lenguaje de interrogación es una herramienta imprescindible para la recuperación de información. En este artículo se ofrece una gramática formal (sintaxis y semántica) para ese tipo de lenguaje. Esta gramática nos permite, por un lado, conocer el lenguaje de interrogación desde un plano sistemático y conceptual y, por otro lado, nos permite obtener ciertos beneficios prácticos de tipo sintáctico y de tipo semántico relacionados con los procesos de recuperación de información.

Palabras clave: lenguaje de interrogación, ecuación de búsqueda, recuperación de información, gramática formal, sintaxis, semántica.

Abstract: The knowledge of query language is an essential tool for information retrieval. This article offers a formal grammar (including syntax and semantics) for this type of language. This grammar provides the systematic and conceptual learning of query language, on the one hand, and the obtention of practical syntactical and semantic benefits related with information retrieval processes, on the other.

Keywords: query language, search statement, information retrieval, formal grammar, syntax, semantics.

1 Introducción

Los usuarios acostumbramos a recurrir a los sistemas de almacenamiento y recuperación de información (o sistemas de información: *SI,* a partir de ahora) con la intención de satisfacer nuestras necesidades informativas. En términos generales, podemos definir este tipo de necesidades como esa clase especial de estados mentales o psicológicos que posee un individuo y cuyo contenido es identificable con un tipo de insatisfacción, curiosidad o disconformidad informativa (1, 2)

Normalmente, materializamos y representamos estos estados mentales mediante el enunciado de un lenguaje natural (castellano, catalán, inglés, etc.). Desgraciadamente, los SI *no* acostumbran a *entender,* por así decirlo, esas peticiones o consultas de información realizadas a partir de los enunciados de un lenguaje natural. Por esta razón, si queremos obtener una respuesta por parte del SI que nos permita satisfacer nuestra necesidad informativa, hemos de transformar esa formulación de manera que el sistema pueda entenderla.

Para cubrir ese objetivo, en una primera fase se realiza un análisis conceptual de la consulta y en la segunda se establece, teniendo en cuenta ese análisis, la traducción de

* Estudis d'Informació i Documentació. Universitat Oberta de Catalunya. Correo-e: mperezgu@campus.uoc.es. http://www.uoc.es

Recibido: Primera versión: 16-9-99. Segunda versión: 24-4-00.

ese enunciado de la lengua natural a un lenguaje determinado accesible para el SI. El lenguaje en cuestión se denomina *lenguaje de interrogación,* y el resultado que obtenemos en este lenguaje mediante la traducción recibe el nombre de *ecuación de búsqueda.*

Finalmente, tras la traducción, se compara (3, 4) la ecuación de búsqueda con las representaciones de los documentos —obtenidas a partir de un proceso de indización (5)—, y se recuperan aquellos documentos cuya representación se ajuste a esa ecuación de búsqueda.

El objetivo principal de todo el proceso es recuperar aquellos documentos que se ajusten de la manera más adecuada a la necesidad de información originaria, es decir, realizar una recuperación de información en la que, en el mejor de los casos, el *ruido* (conjunto de documentos que son recuperados por el SI a partir de la ecuación de búsqueda pero que no se adecuan a la necesidad de información originaria) y el *silencio* (conjunto de documentos que no son recuperados por el SI a partir de la ecuación de búsqueda pero que sí se adecuan a la necesidad de información originaria) sean, hablando en términos conjuntistas, igual al conjunto vacío.

Como se desprende de todo esto, el tema del conocimiento y el manejo del lenguaje de interrogación y el de las ecuaciones de búsqueda se presenta como una herramienta primordial e imprescindible para toda persona, no sólo profesionales, que quiera beneficiarse, en un sentido amplio, de todo el torrente de flujo informativo que nos ofrecen los SI.

Y es que un usuario que se acerque a un SI con la intención de satisfacer una necesidad informativa no podrá extraer todos los beneficios que potencialmente se le brindan a no ser que, entre otras cosas, sepa *entenderse* adecuadamente con el sistema, que se dirija a éste utilizando la *misma* lengua, o dicho en otros términos más técnicos, que conozca de manera adecuada el lenguaje de interrogación y las ecuaciones de búsqueda que lo constituyen.

En definitiva, un conocimiento adecuado de este lenguaje nos permite obtener ciertos beneficios nada despreciables. Por un lado, el más evidente, el dominio de este lenguaje nos otorga un alto grado de autonomía procedimental a la hora de utilizar los SI evitando que recurramos a otra persona, que sí se *entienda* con el sistema, para que nos ayude a realizar la búsqueda y recuperación de información. Por otro lado, el menos evidente, nos permite también tratar y solucionar algunas cuestiones de tipo sintáctico y de tipo semántico estrechamente relacionadas con el uso de este tipo de lenguajes y con la recuperación de información sin tener que recurrir para ello a la poco elegante y (casi siempre) excesivamente cara estrategia del ensayo y el error.

Alimentados por la idea (o esperanza) de que es posible obtener esos beneficios, presentamos aquí un trabajo que tiene como objetivo principal intentar ofrecer una gramática formal (sintaxis y semántica) que nos permita, por un lado, conocer, de una forma adecuada y desde un plano sistemático y conceptual, el lenguaje de interrogación empleado en los procesos de recuperación de información, y, por otro lado, nos habilite para obtener ciertas ventajas de tipo sintáctico y de tipo semántico relacionadas con esos procesos. La idea motriz que fundamenta y articula esos dos objetivos es la de intentar introducir las bases o fundamentos lógicos que puedan servir para que los SI, a partir de mejoras en sus recursos informáticos basadas en nuestra propuesta, tomen decisiones de forma automática respecto a nuestro uso del lenguaje de interrogación y se consiga, de esta manera, simplificar y rentabilizar nuestra interacción con este tipo de sistemas.

Para cubrir ese objetivo, hemos decidido dividir el resto del artículo en tres distintos apartados. Por un lado, en el siguiente apartado (apartado número 2), antes de pasar directamente a abordar el lenguaje que aquí nos ocupa, realizaremos algunas aclaraciones y pondremos de manifiesto algunos aspectos relacionados de forma general con el tema de la gramática y con las propiedades (sintácticas y semánticas) que ésta intenta recoger. Esto nos permitirá, más adelante, ubicar y acotar con mayor precisión nuestra propuesta teórica. A continuación, en el tercer apartado, se intentará mostrar cómo deben ser interpretadas estas ideas cuando hacen referencia a un sistema lingüístico formal como el lenguaje de interrogación que utilizan los SI y se ofrecerá, también, la gramática formal que hemos diseñado pensando en dar cuenta de las propiedades esenciales que caracterizan a ese mismo lenguaje. Y, para finalizar, en el apartado número 4, se intentará poner de manifiesto, desde el ámbito de la sintaxis y de la semántica, una serie de beneficios de tipo conceptual y de tipo práctico que se pueden extraer de la propuesta de gramática formal que se defiende en este trabajo.

2 La gramática como propuesta explicativa de un lenguaje

Dejando al un lado el tema de la pragmática[1], podemos estipular que para poder obtener una gramática de un lenguaje (natural o formal) determinado es necesario contar simultáneamente con una sintaxis y una semántica que expliquen el lenguaje en cuestión (6). O dicho en otros términos, una gramática de un lenguaje concreto nos debe suministrar un conocimiento de las propiedades sintácticas (relaciones que se producen entre los signos de ese lenguaje) y semánticas (relaciones que mantienen esos signos con los objetos que representan) que lo determinan. Los seres humanos (por suerte) tenemos la virtud de utilizar los lenguajes naturales sin tener conciencia explícita, en la mayoría de los casos, de la gramática que los explica. Sin embargo, es interesante describir en qué consisten esas propiedades lingüísticas.

Comencemos por el ámbito de la sintaxis. Los lenguajes se encuentran formados por un conjunto de símbolos que recibe el nombre de «léxico del lenguaje». Algunas combinaciones de esos símbolos dan lugar a ciertas unidades sintácticas mínimas (unidades sintácticas significativas del lenguaje) con las que los usuarios de ese lenguaje pueden llevar a cabo una acción lingüística (transmitir información, expresar una opinión, dar una orden, etc.). La principal propiedad sintáctica que poseen estas unidades y que se intenta recoger a través de una gramática es la de *ser gramatical* o *estar correctamente formada*. Y es que todas las unidades sintácticas significativas del lenguaje son combinaciones de símbolos correctas o gramaticales, pero no todas las combinaciones de símbolos son unidades sintácticas significativas del lenguaje y, por tanto, gramaticales.

Esta propiedad de ser gramatical se caracteriza por ser sistemática y a la vez productiva (7). Se trata de una propiedad sistemática, ya que el conjunto de entidades que poseen la propiedad (el conjunto de combinaciones de elementos del léxico que están bien formadas o son gramaticales, en definitiva) se encuentra determinada por una serie de reglas. La productividad de la propiedad sintáctica consiste, en cambio, en el hecho de que ésta se aplica sobre un conjunto infinito de entidades.

Pasemos ahora al ámbito de la semántica. Para abordar este campo vamos a acotar el territorio de nuestra investigación distinguiendo, de entre todas las unidades sin-

tácticas mínimas (unidades sintácticas significativas del lenguaje) con las que los usuarios pueden llevar a cabo toda clase de acciones lingüísticas (transmitir información, expresar una opinión, dar una orden, etc.), un tipo especial de unidad: los *enunciados.*

Los enunciados son aquellas combinaciones de símbolos utilizadas por parte de los usuarios exclusivamente para realizar ciertos actos lingüísticos específicos: las aseveraciones. O dicho en otros términos, los enunciados son aquellas unidades sintácticas significativas del lenguaje sobre las cuales cabe preguntarnos la verdad o la falsedad de su contenido. La principal propiedad semántica que poseen los enunciados y que también se intenta recoger a través de una gramática es la de *expresar una proposición* o *poseer un contenido susceptible de ser verdadero o falso.*

Como ocurría en el caso de la sintaxis, esta propiedad de expresar una proposición se caracteriza también por ser sistemática y a la vez productiva. Se trata de una propiedad sistemática, ya que la proposición que expresa un enunciado se encuentra siempre determinada por el contenido semántico de las unidades significativas menores que lo conforman. En cambio, la productividad de la propiedad semántica consiste en el hecho de que ésta se aplica sobre un conjunto infinito de entidades.

Las dos propiedades que acabamos de analizar, la sintáctica de la gramaticalidad y la semántica de expresar una proposición, nos permiten explicar por qué los usuarios de un lenguaje poseemos un conocimiento creativo del mismo: porque poseemos la capacidad de producir y entender enunciados que nunca antes nadie había construido. El conocimiento (aunque sea tácito) de estas propiedades y su sistematicidad y productividad es justo lo que nos evita la tarea de tener que aprender de memoria un listado infinito de oraciones y lo que nos habilita para poder utilizar una lengua en términos creativos.

3 Gramática formal para el lenguaje de interrogación

Hasta ahora hemos introducido la idea de que para obtener la gramática de un lenguaje en concreto es necesario describir, principalmente, las dos propiedades que, desde el ámbito de la sintaxis y de la semántica, lo caracterizan: la propiedad sintáctica de la gramaticalidad y la semántica de expresar una proposición.

Ahora cabe preguntarnos cómo podemos proyectar estas ideas sobre un sistema lingüístico formal como el lenguaje de interrogación que utilizan los SI. O dicho en otros términos, ¿cómo podemos obtener una gramática para ese lenguaje?, ¿cómo debemos entender el hecho de que las unidades de este lenguaje ejemplifican esas propiedades? y, además, ¿cómo podemos explicar de forma sistemática y sin ambigüedades ese hecho?

Para contestar a estas preguntas, comencemos primero señalando que, al igual que el resto de los lenguajes, el lenguaje de interrogación (LI, a partir de ahora) que utilizan los SI se encuentra formado por un conjunto de símbolos que podemos identificar como el «léxico del LI». Este conjunto suele estar constituido, a su vez, por un conjunto términos y por los operadores booleanos [AND], [OR] y [NOT][2]. Este lenguaje contiene además dos signos de puntuación: «)» y «(».

Al conjunto de símbolos formado por el léxico del LI y los signos de puntuación lo vamos a denominar «alfabeto del LI». Ningún otro símbolo que no se encuentre comprendido entre los que acabamos de señalar puede considerarse como perteneciente al LI.

Teniendo en cuenta todo esto, pasemos ahora a abordar el ámbito de la sintaxis. Como en el resto de los lenguajes, algunas sucesiones de símbolos del alfabeto del LI dan lugar a las unidades sintácticas significativas del LI.

La principal propiedad sintáctica que poseen estas sucesiones de símbolos es justamente la de *ser una ecuación de búsqueda (ser gramaticales, estar correctamente constituidas)* y diferenciarse, de esta manera, de aquellas sucesiones que no lo son. Y es que todas las ecuaciones de búsqueda son sucesiones gramaticales de símbolos del alfabeto del LI, pero no todas las sucesiones de símbolos de ese alfabeto son ecuaciones de búsqueda y, por tanto, gramaticales.

Como veremos con más detalle un poco más adelante, de la misma manera que en el caso de las oraciones de los lenguajes naturales, esta propiedad de ser una ecuación de búsqueda se caracteriza por ser sistemática ya que el conjunto de entidades que poseen la propiedad (el conjunto de sucesiones de símbolos del alfabeto de LI que son ecuaciones de búsqueda o son gramaticales, en definitiva) se encuentra determinado por una serie de reglas.

La misma propiedad sintáctica de la gramaticalidad se caracteriza también por ser productiva ya que ésta se aplica, como comprobaremos posteriormente, sobre un conjunto infinito de entidades. O dicho en otros términos, debe considerarse como una propiedad productiva ya que el conjunto de sucesiones de símbolos del alfabeto del LI que son ecuaciones de búsqueda es infinito.

Pasemos ahora al ámbito de la semántica. Para abordar este campo vamos a poner de manifiesto, en primer lugar, que los usuarios utilizamos las ecuaciones de búsqueda para llevar cabo una única acción lingüística: realizar una aseveración. En concreto, un usuario del LI, al proponerle una ecuación de búsqueda al SI, lo que pretende es expresar o definir por comprensión[3] cierto conjunto de documentos que quiere que el SI recupere y le permita consultar para satisfacer una necesidad de información original. En este sentido, la principal propiedad semántica que poseen las ecuaciones de búsqueda es la de *expresar o definir un conjunto de documentos*[4].

De la misma manera que ocurría en el caso de los enunciados de los lenguajes naturales, esta propiedad de expresar un conjunto de documentos se caracteriza por ser sistemática ya que, como veremos con más detalle a continuación, el conjunto que expresa o define toda ecuación de búsqueda depende sistemáticamente del contenido semántico de las unidades significativas menores que la constituyen. La misma propiedad semántica se caracteriza también por ser productiva ya que ésta, como comprobaremos posteriormente, se aplica sobre un conjunto infinito de entidades: como existe un conjunto infinito de ecuaciones de búsqueda, el conjunto de entidades que poseen la propiedad semántica, que expresan un conjunto de documentos, es también infinito.

Hasta aquí llegaría nuestra propuesta de cómo deben entenderse las dos principales propiedades (la propiedad sintáctica de ser una ecuación de búsqueda y la semántica de expresar un conjunto de documentos) que caracterizan las unidades sintácticas significativas (las ecuaciones de búsqueda) del LI. Pasemos ahora, en los siguientes subapartados, a intentar explicar de forma sistemática *cuándo* y abundar un poco más sobre el *cómo* esas unidades ejemplifican esas propiedades. Para cubrir este objetivo utilizaremos los recursos conceptuales que nos ofrece la lógica formal (8, 9, 10).

3.1 La sintaxis del lenguaje de interrogación

Antes de introducir la sintaxis del LI, vamos a presentar las convenciones notacionales que utilizaremos en el resto de nuestro trabajo. En este sentido, por un lado, nos auxiliaremos de la letra «T» (completada con subíndices cuando sea necesario) como variable para designar a las palabras o frases de la lengua utilizada en los documentos gestionados por el SI. Por otro lado, utilizaremos las letras mayúsculas «A», «B», «C», etc., como variables de las ecuaciones de búsqueda, y las consonantes minúsculas «d», «f», «g», etc., como variables de los documentos. Y por último, utilizaremos la expresión «[[A]]» como sinónima de la definición por compresión del conjunto de documentos expresada por la ecuación de búsqueda A.

Una vez introducidas las convenciones notacionales, pasemos a presentar nuestra definición recursiva de lo que es una ecuación de búsqueda del LI:

> (I) Una sucesión de símbolos del alfabeto del LI es una ecuación de búsqueda de ese lenguaje si, y sólo si, se encuentra generada por un número finito de aplicaciones de las siguientes reglas:
>
> (a) Si T es una palabra o una frase de la lengua utilizada en los documentos, entonces T es una ecuación de búsqueda del LI
>
> (b) Si A y B son ecuaciones de búsqueda del LI entonces:
>
> (b.i) (A [AND] B) es también una ecuación de búsqueda del LI
>
> (b.ii) (A [OR] B) es también una ecuación de búsqueda del LI
>
> (b.iii) (A [NOT] B) es también una ecuación de búsqueda del LI

Esta definición contenida en (I) nos genera el conjunto de todas las posibles sucesiones de símbolos del alfabeto de LI que poseen la propiedad sintáctica de ser una ecuación de búsqueda de ese mismo lenguaje. Veamos, ahora, en qué sentido podemos justificar, a partir de la definición, el hecho de que esta propiedad se caracterice por ser sistemática y a la vez productiva.

Por un lado, el hecho de que la propiedad de ser una ecuación de búsqueda posea la característica de la sistematicidad se justifica de una manera evidente: como se desprende de (I), que una sucesión de símbolos sea o no sea una ecuación de búsqueda depende de que ésta cumpla o no algunas de las reglas descritas por (a) o (b). Así, por ejemplo, la sucesión de símbolos del alfabeto ((Política [AND] Economía) [NOT] Siglo XIX) debe ser considerada como una ecuación de búsqueda (suponiendo, claro está, que el lenguaje de algunos documentos gestionados por el SI sea el castellano) ya que se obtiene a partir de tres aplicaciones de las reglas en cuestión:

> 1. Política, Economía y Siglo XIX son ecuaciones de búsqueda. (Se obtienen por la regla (a)).
> 2. (Política [AND] Economía) es una ecuación de búsqueda. (Se obtiene a partir de 1 y aplicando la regla (b.i)).
> 3. ((Política [AND] Economía) [NOT] Siglo XIX)) es una ecuación de bús-

queda, como queríamos demostrar. (Se obtiene a partir de 1, 2 y aplicando la regla (b.iii)).

En cambio, la sucesión de símbolos (Política [AND] Economía [NOT] Siglo XIX) no podrá ser considerada como una ecuación de búsqueda del LI ya que no es posible generarla (le faltan algunos signos de puntuación) aplicando las reglas contenidas en la definición (I).

Por otro lado, la característica de la productividad que define la propiedad de ser una ecuación de búsqueda del LI se justifica principalmente por el carácter recursivo de la regla (b) de la definición en cuestión. Tal y como están introducidas las subreglas (b.i), (b.ii) y (b.iii), permiten que éstas puedan aplicarse de nuevo a los productos obtenidos de la aplicación de las propias reglas, generando de esta manera un conjunto infinito de ecuaciones de búsqueda de LI.

En este sentido, por ejemplo, si obtenemos la ecuación de búsqueda (Política [AND] Economía) a partir de los términos *Política y Economía* y la regla (b.i), podemos obtener también otra nueva ecuación si combinamos por partida doble esa ecuación originaria mediante la misma (u otra) regla recursiva: ((Política [AND] Economía) [AND] (Política [AND] Economía)). De nuevo, podemos volver a aplicar el mismo procedimiento progresivamente e iremos formando un conjunto infinito de sucesiones del alfabeto del LI que pueden considerarse como ecuaciones de búsqueda de ese lenguaje.

3.2 La semántica del lenguaje de interrogación

Como ya hemos apuntado, la principal propiedad semántica que caracterizan las ecuaciones de búsqueda del LI es la de expresar, representar o identificar por comprensión un conjunto de documentos. Pero antes de pasar a intentar definir esa propiedad y justificar sus características, es importante poner de manifiesto un aspecto que puede ayudarnos a entender el contenido de este apartado.

El aspecto en cuestión es el siguiente: teniendo en cuenta que en la fase de indización se asocia cada documento con una serie de términos (o descriptores) que lo representan, podemos identificar en clave conjuntista cada uno de los documentos gestionados por el SI, es decir, podemos identificar cada documento como el conjunto de esos términos que se obtiene a partir de su indización. Así, por ejemplo, si tras el proceso de indización del documento d se obtiene que T_1, T_2 y T_3 son sus términos representativos (también denominados «términos de indización»), podemos identificar d de la siguiente manera: $d = \{T_1, T_2 \text{ y } T_3\}$.

Con esta aclaración en la mano, ya podemos introducir nuestra definición de la propiedad semántica que ejemplifican las ecuaciones de búsqueda del LI:

> (II) La definición por comprensión de un conjunto de documentos expresada por una ecuación de búsqueda del LI ([[ecuación de búsqueda]]) se obtiene a partir de un número finito de aplicaciones de las siguientes reglas:
>
> (a) Si T es una palabra o una frase de la lengua utilizada en los documentos gestionados por el SI, entonces:
>
> $$[[T]] = \{d: T \in d\}$$

(b) Si A y B son ecuaciones de búsqueda del LI entonces:
(b.i) [[(A [AND] B)]] = {d: d ∈ [[A]] y d ∈ [[B]]}
(b.ii) [[(A [OR] B)]] = {d: d ∈ [[A]] o d ∈ [[B]]}
(b.iii) [[(A [NOT] B)]] = {d: d ∈ [[A]] y d ∈ [[B]]}

La regla (a) nos indica que el conjunto de documentos expresado por la ecuación de búsqueda T se encuentra formado por todos aquellos documentos d que incluyan el término T entre los términos que lo representan.

La regla (b.i) nos dice que el conjunto expresado por la ecuación de búsqueda (A [AND] B) está constituido por aquellos documentos d que pertenecen al conjunto representado por la ecuación A y que pertenecen a la vez al conjunto expresado por la ecuación B.

De forma general, utilizaremos la expresión «d ∈ [[A]]» como sinónima de las expresiones «d pertenece al conjunto representado por la ecuación A», «d es un documento verdadero respecto a la ecuación A», «d satisface la ecuación A» y «d es lógicamente relevante según la ecuación A». De la misma manera, utilizaremos la expresión «d ∉ [[B]]» como sinónima de las expresiones «d no pertenece al conjunto representado por la ecuación A», «d no es un documento verdadero respecto a la ecuación A», «d no satisface la ecuación A» y «d no es lógicamente relevante según la ecuación A».

La regla (b.ii), en cambio, nos indica que el conjunto representado por la ecuación de búsqueda (A [OR] B) lo forman aquellos documentos d que pertenecen al conjunto expresado por la ecuación A o al conjunto expresado por la ecuación B. Hay que señalar que en este contexto no se utiliza la disyunción «o» en un sentido excluyente. Esto significa que en el conjunto definido por esa ecuación se encuentran los documentos que pertenecen a [[A]], los que pertenecen a [[B]] y los que pertenecen simultáneamente a [[A]] y a [[B]].

Por ultimo, la regla (b.iii) nos señala que el conjunto expresado por la ecuación de búsqueda (A [NOT] B) se encuentra constituido por aquellos documentos d que pertenecen al conjunto representado por la ecuación A y que en cambio no pertenecen al conjunto expresado por la ecuación B. Es importante señalar que mientras que los operadores [AND] y [OR] son conmutativos ya que [[A [AND] B]] = [[B [AND] A]] y [[A [OR] B]] = [[B [OR] A]], el operador [NOT] no lo es ya que [[A [NOT] B]] ≠ [[B [NOT] A]], como se desprende de las representaciones gráficas que se introducen a continuación.

De manera simultánea, podemos obtener una representación gráfica mediante diagramas de las reglas (a), (b.i), (b.ii) y (b.iii) a partir de las figuras 1, 2, 3 y 4, respectivamente. El área sombreada se corresponde con el conjunto de documentos definido por la ecuación de búsqueda.

Figura 1
Representación gráfica de [[T]]

Figura 2
Representación gráfica de [[(A[AND]B)]]

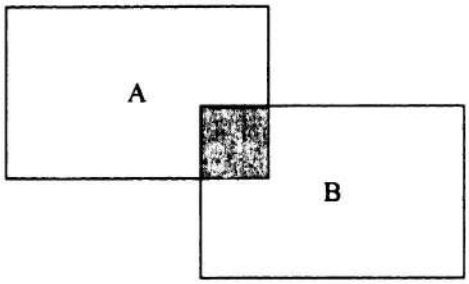


Figura 3
Representación gráfica de [[(A[OR]B)]]

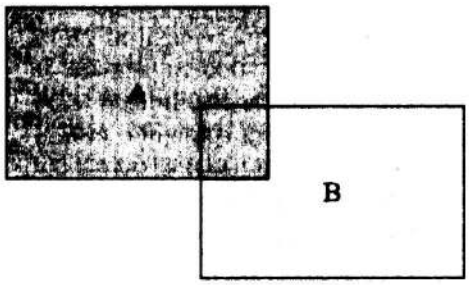
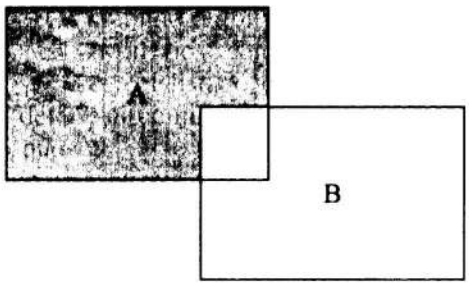


Figura 4
Representación gráfica de [[(A[NOT]B)]]



Veamos, ahora, en qué sentido podemos justificar, a partir de la definición contenida en (II) el hecho de que esta propiedad semántica de expresar un conjunto de documentos se caracterice por ser sistemática y a la vez productiva.

La sistematicidad de la propiedad semántica se justifica por el hecho de que hemos establecido un fuerte paralelismo entre las reglas contenidas en (I) y las presentadas en (II): a cada una de las reglas sintácticas le hemos hecho corresponder una regla semántica. En definitiva, se han aprovechado las reglas sintácticas de formación de las ecuaciones de búsqueda (las contenidas en (I)) para basar o fundamentar sobre éstas la definición de la propiedad semántica que se presenta en (II). De esta manera, se ha conseguido mostrar cómo el conjunto de documentos que expresa o define por comprensión toda ecuación de búsqueda depende sistemáticamente del contenido semántico de las unidades significativas menores que la constituyen.

Así, por ejemplo, si recuperamos la ecuación introducida en el apartado anterior, ((Política [AND] Economía) [NOT] Siglo XIX), y aplicamos la definición contenida en (II) vemos claramente que el conjunto de documentos expresado por ésta depende de los conjuntos expresados por las ecuaciones mínimas que la componen (Política, Economía, Siglo XIX) y por las ecuaciones que obtenemos al combinar éstas mediante los operadores [AND] y [NOT].

Esta simetría entre las reglas sintácticas y las semánticas nos permite, también, poner de manifiesto uno de los rasgos básicos que presentan los lenguajes en general: los cambios sintácticos que se introducen en las expresiones acostumbran a provocar además cambios semánticos.

Para ilustrar esta idea sólo tenemos que elegir una ecuación de búsqueda, como por ejemplo ((Política [AND] Economía) [OR] Siglo XIX), e introducir unos cambios sintácticos mediante la variación de sus signos de puntuación, de sus paréntesis, hasta obtener una segunda ecuación de búsqueda, como por ejemplo: (Política [AND] (Economía [OR] Siglo XIX)).

Ahora, si aplicamos la definición contenida en (II) obtenemos que esos mínimos cambios sintácticos han provocado también cambios semánticos[5], mientras que la primera ecuación de búsqueda representa el conjunto de documentos d tal que d pertenece a [[(Política [AND] Economía)]] o pertenece a [[Siglo XIX]][6], la segunda representa el conjunto de documentos d tales que d pertenece simultáneamente a [[Política]] y a [[(Economía [OR] Siglo XIX]][7]. Esta diferencia se hace más evidente cuando comprobamos, por ejemplo, que un documento que contenga el término *Siglo XIX* pero que no contenga los términos *Política* y *Economía* pertenecerá al conjunto de documentos representados por la primera ecuación y en cambio no estará incluido en el conjunto de documentos representados por la segunda.

La diferencia entre estos dos conjuntos de documentos se puede apreciar también si nos remitimos a la representación gráfica mediante diagramas de cada una de esas dos ecuaciones de búsqueda realizadas respectivamente en las figuras 5 y 6. De nuevo, las áreas sombreadas se corresponden con los conjuntos representados.

Pasemos ahora a mostrar cómo la misma propiedad semántica de representar conjuntos de documentos se caracteriza también por ser productiva. Este rasgo se

Figura 5
Representación gráfica de [[((Política [AND] Economía) [OR] Siglo XIX)]]

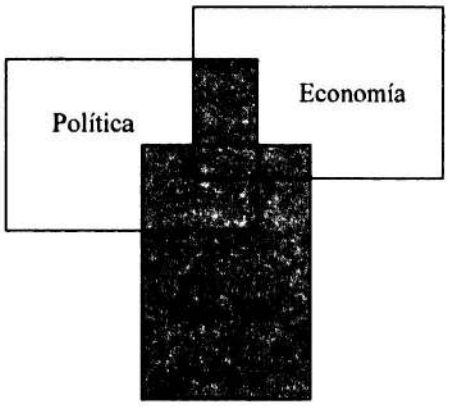
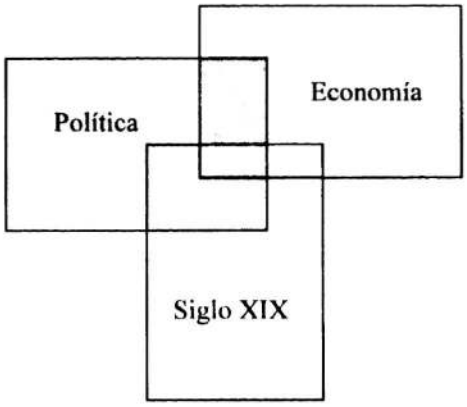


Figura 6
Representación gráfica de [[(Política [AND] (Economía [OR] Siglo XIX))]]



justifica, de nuevo, por la especial naturaleza de la regla semántica (b) de la definición (II).

Esta regla semántica, al estar fundamentada sobre la regla de formación sintáctica (b) de (I) y su carácter recursivo, es susceptible de poder aplicarse sobre un conjunto infinito de ecuaciones de búsqueda. Y es que, al quedar justificado —como vimos en el apartado anterior— que se puede generar un conjunto infinito de estas ecuaciones a partir de la regla (b) de (I), queda también justificado, por la relación de simetría o fundamentación entre semántica y sintaxis, que es posible obtener, mediante la regla (b) de (II), el conjunto de documentos representado por cada una de esas ecuaciones.

O dicho en otros términos: a partir de esa simetría, se podrá sostener que cualquier sucesión de signos del alfabeto del LI que podamos imaginar, por extensa que sea, y que se encuentre bien formada, es decir, que sea una ecuación de búsqueda, representará un conjunto de documentos. No existe *ninguna* sucesión de símbolos de ese alfabeto que sea una ecuación y que en cambio *no* represente algún conjunto de documentos.

Así, por ejemplo, aunque a partir de la ecuación de búsqueda (Política [AND] Economía) sea posible obtener un conjunto infinito de sucesiones del alfabeto del LI que pueden considerarse como ecuaciones de búsqueda del LI —como ya vimos en el apartado anterior—, será también posible obtener el conjunto de documentos representados por cada una de éstas a partir de [[Política]] y [[Economía]] (del contenido de las ecuaciones de búsqueda simples Política y Economía) y de las reglas semánticas relacionadas con cada uno de los operadores implicados.

3.3 Apéndice: los operadores de proximidad

Hasta el momento hemos limitado el conjunto de operadores pertenecientes al alfabeto de LI a los operadores booleanos [AND], [OR] y [NOT]. El motivo de esta limitación responde principalmente a dos criterios. Por un lado, hemos seguido la inercia que se comprueba dentro de la literatura especializada y que destaca la importancia de este tipo de operadores frente al resto. Y, por otro lado, hemos decidido acotar el objeto de estudio para conseguir una mayor claridad expositiva.

Sin embargo, en el LI acostumbran a utilizarse también otros tipos de operadores.

Entre estos hay que destacar los denominados «operadores de proximidad»: [·], [near] y [n]. Aunque no vamos a abundar excesivamente sobre éstos, en este apéndice vamos a proponer cómo podrían tratarse esos operadores tanto desde el punto de vista sintáctico como desde el semántico, completando de esta manera nuestras definiciones de ecuación de búsqueda y de conjunto de documentos representado por una ecuación de búsqueda.

Como tratamiento sintáctico de los operadores de proximidad proponemos la siguiente regla (c) que podría añadirse y completar de esta manera la definición de ecuación de búsqueda contenida en (I):

> (c) Si T_1, T_2 son palabras o frases de la lengua utilizada en los documentos gestionados por el SI, entonces:
>
> (c.i) (T1 [·] T2) es una ecuación de búsqueda del LI
>
> (c.ii) (T1 [near] T2) es una ecuación de búsqueda del LI
>
> (c.iii) (T1 [n] T2) es una ecuación de búsqueda del LI

Así, por ejemplo, si *Política* y *Economía* son palabras utilizadas en los documentos gestionados por el SI entonces las sucesiones de símbolos (Política [·] Economía), (Política [near] Economía) y (Política [n] Economía) deben ser consideradas como ecuaciones de búsqueda del LI.

Es importante señalar que la regla contenida en (c) no es de tipo recursivo. Es decir, los operadores de proximidad no se pueden aplicar sobre los resultados que se obtienen a partir de los mismos. Esto significa que sólo se pueden emplear entre términos y no entre otro tipo de ecuaciones de búsqueda más complejas (compuestas a su vez por ecuaciones), aunque sí pueden formar parte de ecuaciones de búsqueda más complejas al combinarse con otras a partir de la regla (b) de (I), como por ejemplo en ((Política [·] Economía) [AND] Siglo XIX).

Como tratamiento semántico de este tipo de operadores proponemos la siguiente regla (c') que podría añadirse y completar de esta manera la definición del conjunto de documentos representado por una ecuación de búsqueda contenida en (II):

> (c') Si T_1 y T_2 son palabras o frases de la lengua utilizada en los documentos gestionados por el SI, entonces:
>
> (c'.i) $[[(T_1\ [\cdot]\ T_2)]] = \{d: d \in [[(T_1\ [AND]\ T_2)]]$ y, además, en d, T_1 y T_2 aparecen en la misma frase$\}$
>
> (c'.ii) $[[(T_1\ [near]\ T_2)]] = \{d: d \in [[(T_1\ [AND]\ T_2)]]$ y, además, en d, T_1 y T_2 aparecen en el mismo párrafo$\}$
>
> (c'.iii) $[[(T_1\ [n]\ T_2)]] = \{d: d \in [[(T_1\ [AND]\ T_2)]]$ y, además, en d, T_1 y T_2 no aparecen separados por un número máximo (n) de palabras$\}$

La regla (c'.i) nos dice que el conjunto expresado por la ecuación de búsqueda (T_1 [·] T_2) está constituido por aquellos documentos d que incluyen T_1 y T_2 entre sus términos de indización y que, a la vez, en d aparecen esos dos términos dentro de una misma frase.

La regla (c'.ii), en cambio, nos indica que el conjunto representado por la ecuación de búsqueda (T_1 [near] T_2) lo forman aquellos documentos d que cuentan con T_1 y T_2 cómo términos de indización y que, a la vez, en d esos dos términos están ubicados dentro de un mismo párrafo.

Por ultimo, la regla (c'.iii) nos señala que el conjunto expresado por la ecuación de búsqueda (T_1 [n] T_2) se encuentra constituido por aquellos documentos d que contemplan T_1 y T_2 como términos que se obtienen a partir de su indización y que, a la vez, en d esos dos términos no aparecen separados por un número (n) máximo determinado de palabras.

4 Los beneficios explicativos de la gramática propuesta

En este último apartado vamos a intentar poner de manifiesto, de forma resumida, algunos de los beneficios que se pueden extraer de nuestra propuesta de gramática formal para el LI que acabamos de introducir. Para presentarlos, vamos a utilizar la estrategia de distinguir dos grupos diferenciados: el que formarían los beneficios conceptuales, por un lado, y el constituido por los beneficios de tipo práctico, por otro. Cada uno de estos grupos admite en su seno, a su vez, la distinción entre aquellos que se encuentran relacionados con el campo de la sintaxis y aquellos que se derivan del ámbito de la semántica.

Comencemos por los beneficios de tipo conceptual. Dentro de este grupo se van a intentar recoger los resultados teóricos básicos de nuestra propuesta. En este sentido, las ideas que vamos a introducir en los dos siguientes párrafos pueden ser entendidas como una versión resumida de los principales puntos expuestos hasta el momento en este trabajo.

En el campo de la sintaxis hemos obtenido una serie de claros beneficios conceptuales. Por un lado, hemos identificado *cuál* es la propiedad sintáctica de las unidades significativas del LI. Esa propiedad, en concreto, es la de *ser una ecuación de búsqueda.* Por otro lado hemos ofrecido, en (I), una definición de ecuación de búsqueda. En este contexto, entendemos por definición una serie de condiciones suficientes y necesarias que debe cumplir toda sucesión de símbolos del alfabeto del LI para que pueda ser considerada como una ecuación de búsqueda de ese mismo lenguaje. Por último, hemos mostrado en qué sentido se puede entender la sistematicidad y la productividad que caracterizan a esta propiedad sintáctica.

En el ámbito de la semántica también hemos podido extraer una serie de beneficios teóricos. Por un lado, hemos establecido *cuál* es la principal propiedad semántica de las ecuaciones de búsqueda o unidades significativas del LI. Esa propiedad ha sido identificada como la de *representar un determinado conjunto de documentos.* Por otro lado, se ha introducido, en (II), una definición de esta propiedad. En este caso, por definición entendemos una serie de mecanismos formales que nos permiten determinar el contenido semántico (cuál es el conjunto de documentos representado) de cualquier sucesión de símbolos que sea una ecuación de búsqueda del LI. Por último, hemos señalado cómo esa misma propiedad semántica se caracteriza por el hecho de ser simultáneamente productiva y sistemática.

Abandonemos los beneficios conceptuales y concentremos nuestra atención, ahora, sobre los provechos de tipo práctico que se pueden extraer de nuestra propuesta teórico-formal. Dentro de este grupo vamos a intentar poner de manifiesto aquellas utilidades y ventajas que podemos obtener de esa propuesta cuando participamos en procesos o praxis reales de recuperación de información a través de algún SI.

Comencemos por los beneficios prácticos relacionados con el ámbito de la sintaxis. El primero de estos beneficios es evidente: tal y como se ha introducido la definición contenida en (I) nos encontramos en disposición de poder decidir sin el menor tipo de ambigüedad si una sucesión de símbolos del alfabeto del LI, por larga y complicada que nos parezca, es o no realmente una ecuación de búsqueda de ese mismo lenguaje. Ilustremos esta posibilidad a través de una serie de ejemplos.

Ejemplo 1. ¿La sucesión de símbolos del alfabeto de LI ((Política [AND] Economía) [OR] (Siglo XIX [NOT] Emigración)) puede ser considerada como una genuina ecuación de búsqueda de ese lenguaje? Respuesta: Sí.

Justificación:

1. Política, Economía, Siglo XIX y Emigración son ecuaciones de búsqueda. (Se obtienen por la regla (a) de (I)).
2. (Política [AND] Economía) es una ecuación de búsqueda. (Se obtiene a partir de 1 aplicando la regla (b.i)).
3. (Siglo XIX [NOT] Emigración) es una ecuación de búsqueda. (Se obtiene a partir de 1 aplicando la regla (b.iii) de (I)).
4. ((Política [AND] Economía) [OR] (Siglo XIX [NOT] Emigración)) es una ecuación de búsqueda, como queríamos demostrar. (Se obtiene a partir de 2 y 3 aplicando la regla (b.ii) de (I)).

Ejemplo 2. ¿La sucesión de símbolos del alfabeto de LI (Política [AND] Economía) [OR] (Siglo XIX [NOT] Emigración)) puede ser considerada como una genuina ecuación de búsqueda de ese lenguaje? Respuesta: No.

Justificación:

1. Política, Economía, Siglo XIX y Emigración son ecuaciones de búsqueda. (Se obtienen por la regla (a) de (I)).
2. (Política [AND] Economía) es una ecuación de búsqueda. (Se obtiene a partir de 1 aplicando la regla (b.i)).
3. (Siglo XIX [NOT] Emigración) es una ecuación de búsqueda. (Se obtiene a partir de 1 aplicando la regla (b.iii) de (I)).
4. (Política [AND] Economía) [OR] (Siglo XIX [NOT] Emigración)) no es una ecuación de búsqueda, como queríamos justificar. (No existe ninguna regla en (I) que genere esta sucesión de símbolos a partir de 1, 2 o 3).

Ejemplo 3. ¿La sucesión de símbolos del alfabeto de LI (Política [AND] ((Economía [OR] Siglo XIX) [NOT] Emigración)) puede ser considerada como una genuina ecuación de búsqueda de ese lenguaje? Respuesta: Sí.

Justificación:

1. Política, Economía, Siglo XIX y Emigración son ecuaciones de búsqueda. (Se obtienen por la regla (a) de (I)).
2. (Economía [OR] Siglo XIX) es una ecuación de búsqueda. (Se obtiene a partir de 1 aplicando la regla (b.ii)).

3. ((Economía [OR] Siglo XIX) [NOT] Emigración) es una ecuación de búsqueda. (Se obtiene a partir de 1 y 2 aplicando la regla (b.iii) de (I)).
4. (Política [AND] ((Economía [OR] Siglo XIX) [NOT] Emigración)) es una ecuación de búsqueda, como queríamos demostrar. (Se obtiene a partir de 1 y 3 aplicando la regla (b.i) de (I)).

El segundo de los beneficios sintácticos se obtiene cuando aprovechamos esa capacidad discriminatoria que nos ofrece la definición contenida en (I) en un tipo especial de circunstancias. En concreto, hay que señalar que en algunas situaciones, cuando proponemos una sucesión de símbolos del LI a un SI para que nos recupere un conjunto de documentos que en teoría puede satisfacer nuestra necesidad de información, existe la posibilidad de que el sistema nos responda que no existen documentos que se adecuen a las condiciones representadas por esa sucesión de símbolos.

En este tipo de situaciones siempre nos queda una duda: no podemos saber si esa respuesta del sistema se justifica porque no existe realmente un conjunto de documentos que se corresponda con esa ecuación de búsqueda o porque la sucesión de símbolos que hemos propuesto al sistema no es una genuina ecuación de búsqueda y por tanto no representa ningún conjunto de documentos susceptibles de ser recuperados. Si la respuesta del sistema se justifica por la primera razón y la ecuación de búsqueda en cuestión es la traducción correcta del enunciado del lenguaje natural que representa nuestra necesidad de información, nos veremos obligados a buscar otro SI que pueda servirnos para satisfacer realmente nuestra necesidad de información. En cambio, si la respuesta se justifica por la segunda razón, nos veremos obligados a reformular la sucesión de símbolos hasta convertirla en una genuina ecuación de búsqueda.

La definición contenida en (I) nos permite resolver esa duda sin ningún tipo de ambigüedad: nos habilita para discriminar cuál ha sido, de las dos, la razón que justifica la respuesta del sistema.

En este sentido, cuando nos veamos involucrados en una situación de este tipo, sólo tendremos que decidir mediante (I) si la sucesión de símbolos en cuestión es o no realmente una genuina ecuación de búsqueda.

Si lo es y no ha habido problemas en la traducción, abandonaremos el SI porque éste es incapaz de indicarnos documentos que pueden satisfacer nuestra necesidad informativa (quizá el SI gestiona documentos alejados del tema que nos interesa). Esto es lo que ocurriría, por ejemplo, si tras traducir correctamente e introducir las ecuaciones de búsqueda presentadas en los enunciados de los anteriores ejemplos 1 y 3, el sistema nos respondiese que no existen documentos que respondan a esas condiciones.

Si no es una genuina ecuación de búsqueda, tendremos que replantear el hecho de que esa sucesión de símbolos pueda ser considerada como la traducción adecuada al LI de nuestra necesidad de información. Esto es lo que ocurriría, por ejemplo, si el sistema nos ofreciese la misma respuesta negativa tras introducir la sucesión de símbolos presentada en el enunciado del ejemplo 2.

Esta posibilidad de discriminación que nos ofrece (I) no debe ser entendida como una cuestión baladí, ni tampoco como un brindis al sol: además de resolver la duda anteriormente planteada, con esta capacidad de decisión podemos alcanzar un considerable ahorro de tiempo y dinero a la hora de intentar recuperar una serie de documentos gestionados por un SI. Ahorraremos un tiempo considerable porque no reali-

zaremos más consultas que aquéllas en las que introduzcamos genuinas ecuaciones de búsqueda (aquéllas que nos permitan decidir si el SI realmente nos puede ayudar o no a satisfacer nuestra necesidad de información); y, por tanto, nos ahorraremos cierta cantidad de dinero ya que se verá reducido el número de consultas que tendremos que pagar (suponiendo que nos cobren por cada una de las consultas realizadas).

Pasemos, ahora, a mostrar el beneficio de tipo práctico que obtenemos desde el ámbito de la semántica. Como ya indicamos al inicio de este trabajo, el objetivo principal que debe perseguir todo usuario que decide consultar un SI para satisfacer una necesidad informativa consiste en que el sistema le recupere, a partir de una determinada ecuación de búsqueda y con un ruido y un silencio igual al conjunto vacío, un conjunto de documentos adecuados a esa necesidad.

Evidentemente, una de las variables fundamentales que intervienen en el éxito o el fracaso de la consulta se encuentra en el hecho de que la ecuación de búsqueda se adecue correctamente a esa necesidad de información. O dicho en otros términos: que la ecuación represente adecuadamente el conjunto de documentos en el que el usuario del SI está interesado. Los usuarios no familiarizados con los SI acostumbran a realizar este proceso recurriendo a sus intuiciones y empleando, en la mayoría de los casos, la poco rentable estrategia del ensayo y el error: a partir de sus intuiciones (o como buenamente pueden) proponen al sistema una ecuación de búsqueda y comprueban más tarde, una vez realizada la recuperación, si los documentos recuperados se corresponden o no con la necesidad de información. Si se corresponden abandonan la búsqueda, pero si no es el caso vuelven a repetir la operación hasta que consiguen que se correspondan (o desisten en su intento por aburrimiento).

La definición contenida en (II) nos puede ayudar a garantizar el control de esa variable para realizar con éxito la consulta y despreciar, de esta manera, la estrategia del ensayo y el error. Para conseguir esto, por un lado, hemos de formular la necesidad de información en términos conjuntistas, es decir, identificar cuál es el conjunto de documentos que requiere esa necesidad para verse satisfecha. Por otro lado, debemos proponer las ecuaciones de búsqueda que creamos que son las candidatas más adecuadas para representar esa necesidad. Por último, aplicando la definición (II) podemos concluir de forma exacta cuál es el conjunto de documentos que representa cada una de éstas y decidir, utilizando como criterio ese resultado, cuál es la ecuación que mejor representa la necesidad de información original, es decir, cuál es la ecuación cuyo conjunto de documentos que representa es igual al conjunto que se obtiene de la lectura de la necesidad de información en términos conjuntistas. Una vez elegida la ecuación en cuestión ya podemos proponérsela al SI[8].

Para ilustrar el beneficio semántico que extraemos de nuestra propuesta sólo hay que recurrir a un sencillo ejemplo. Imaginemos que un usuario se acerca a un SI para conseguir documentos que le permitan profundizar en la relación, en el ámbito mundial, entre la emigración, por un lado, y la política o la economía, por otro, excluyendo lo ocurrido en Francia. El primero de nuestros pasos es conseguir efectuar una lectura conjuntista de esa necesidad de información. En este caso podemos decir que el usuario quiere obtener documentos que contengan siempre el término *Emigración,* que contengan indistintamente el término *Economía* o el término *Política,* pero que sin embargo no incluyan el término *Francia.* Si recurrimos a la simbología de la Teoría de Conjuntos obtenemos que esa necesidad puede identificarse con el siguiente conjunto: {d: d ∈ [[Emigración]], y d ∈ [[Economía]] o d ∈ [[Política)]], pero d ∉ [[Francia]]}.

A continuación, a partir de los términos que intervienen y los posibles operadores, vamos a proponer dos ecuaciones de búsqueda como candidatas para representar esa necesidad de información: la ecuación A, (((Emigración [AND] Economía) [OR] Política) [NOT] Francia), y la ecuación B, ((Emigración [AND] (Economía [OR] Política)) [NOT] Francia). La duda que se nos plantea ahora es: ¿cuál de las dos ecuaciones, A o B, es la que representa la necesidad, si es el caso? Para responder a esta pregunta vamos a averiguar, utilizando para ello la definición contenida en (II), el conjunto que representa cada una de ellas.

De esta manera, la ecuación A, (((Emigración [AND] Economía) [OR] Política) [NOT] Francia), representa el siguiente conjunto de documentos:

1. [[A]] = {d: d ∈ [[((Emigración [AND] Economía) [OR] Política)]], pero d ∉ [[Francia]]} (Se obtiene a partir de la regla (b.iii) de (II)).
2. [[A]] = {d: d ∈ [[((Emigración [AND] Economía)]] o d ∈ [[Política)]], pero d ∉ [[Francia]]} (Se obtiene a partir de 1 aplicando la regla (b.ii) de (II)).
3. [[A]] = {d: d ∈ [[Emigración]] y d ∈ [[Economía]], o d ∈ [[Política)]], pero d ∉ [[Francia]]} (Se obtiene a partir de 2 aplicando la regla (b.i) de (II)).

De forma análoga, la ecuación B, ((Emigración [AND] (Economía [OR] Política)) [NOT] Francia), representa el siguiente conjunto de documentos:

1. [[B]] = {d: d ∈ [[(Emigración [AND] (Economía [OR] Política))]], pero d ∉ [[Francia]]} (Se obtiene a partir de la regla (b.iii) de (II)).
2. [[B]] = {d: d ∈ [[((Emigración]], d ∈ [[(Economía [OR] Política)]], pero d ∉ [[Francia]]} (Se obtiene a partir de 1 aplicando la regla (b.i) de (II)).
3. [[B]] = {d: d ∈ [[Emigración]], y d ∈ [[Economía]] o d ∈ [[Política)]], pero d ∉ [[Francia]]} (Se obtiene a partir de 2 aplicando la regla (b.ii) de (II)).

Una vez obtenidos los conjuntos representados por cada una de las ecuaciones candidatas, concluimos que la ecuación B es la que mejor representa la necesidad de información original. O dicho de otra manera, B es la ecuación cuyo conjunto de documentos que representa es igual al conjunto que se obtiene de la lectura de la necesidad de información en términos conjuntistas. Ahora ya podemos proponer B al SI y esperar que recupere los documentos para satisfacer esa necesidad anteriormente descrita.

De nuevo, hemos de señalar que, al igual que ocurría en el segundo de los beneficios sintácticos, este provecho que extraemos de la definición (II) no es en absoluto despreciable: al ayudarnos en el diseño de la ecuación más adecuada a la necesidad informativa, nos permite evitar la poco rentable estrategia del ensayo y el error y alcanzar, de esta manera, un considerable ahorro de tiempo y dinero a la hora de intentar recuperar una serie de documentos gestionados por el SI. Ahorraremos un tiempo considerable porque no realizaremos más consultas que aquéllas en las que introduzcamos la ecuación de búsqueda que consideremos como la más adecuada; y, por tanto, nos ahorraremos cierta cantidad de dinero ya que se verá reducido el nú-

mero de consultas que tendremos que pagar (suponiendo, de nuevo, que nos cobren por cada una de las consultas realizadas).

Para finalizar, sólo nos queda mostrar la idea que ha alimentado, aunque sea de una forma indirecta, el desarrollo de todo este trabajo y que se encuentra estrechamente relacionada con los beneficios prácticos de tipo sintáctico y semántico que acabamos de describir. La idea en cuestión es que a partir de los beneficios conceptuales que se extraen de nuestra propuesta (la definición de la principal propiedad sintáctica y semántica de las unidades significativas del lenguaje de interrogación) hemos intentado también introducir y ofrecer las bases o fundamentos lógicos que pueden ayudar, al menos en parte, a que, en un segundo movimiento, los programadores desarrollen —tarea que ya no nos corresponde— los recursos adecuados para completar y mejorar los SI y el comportamiento que éstos presentan.

Esa mejoría de los SI consistiría en rediseñar, a partir de las ideas aquí presentadas, los programas informáticos que los sustentan para que éstos tomaran decisiones de forma automática respecto a nuestro uso del lenguaje de interrogación dentro del sistema.

Hasta el momento los beneficios prácticos que hemos descrito se pueden obtener, como hemos visto, si el *usuario* del SI aplica de una forma directa las definiciones contenidas en (I) y (II) a la hora de utilizar el lenguaje de interrogación. En cambio, nuestra propuesta pretende sentar las bases o abrir la puerta para que se dé un paso más: que sea el *propio SI* el que de forma automatizada tome ciertas decisiones de tipo sintáctico y de tipo semántico para que el usuario pueda hacerse con esos beneficios de una forma más sencilla, sin tener que aplicar directamente las definiciones (I) y (II), en su interacción con el sistema.

De esta manera, en el plano sintáctico, a partir del diseño informático fundamentado en la definición (I), se pueden conseguir dos cosas importantes. Por un lado, que cada vez que propongamos al SI una sucesión de símbolos del alfabeto del LI, éste nos conteste de una forma automática y sin ningún tipo de ambigüedad si esa sucesión, por larga y complicada que nos parezca, es o no realmente una ecuación de búsqueda de ese mismo lenguaje. Y, por otro lado, que aprovechando ese beneficio podamos discriminar si, cuando un SI nos responde que no existen documentos que se adecuen a las condiciones representadas por una sucesión de símbolos del LI, esa respuesta del sistema se fundamenta en el hecho de que no existe realmente un conjunto de documentos que se corresponda con esa ecuación de búsqueda o en el hecho de que la sucesión de símbolos propuesta no es una genuina ecuación de búsqueda.

En la misma línea, pero en el ámbito semántico, a partir del diseño informático fundamentado en la definición (II), se puede conseguir también que el SI nos ofrezca de forma automática cuál es el conjunto, definido por comprensión, de documentos que recuperaríamos si éste realizase una búsqueda a partir de una ecuación determinada. De esta manera, aprovechando ese recurso que nos ofrece el sistema, podemos elegir, de entre un conjunto de ecuaciones de búsqueda alternativas, cuál es la que se adecua correctamente a una necesidad concreta de información, o dicho de otra manera, cuál es la ecuación que representa adecuadamente el conjunto de documentos en el que el usuario del SI está interesado. Y todo ello, claro está, sin vernos obligados a provocar que el sistema recupere los documentos a partir de cada una de las ecuaciones alternativas, es decir, sin recurrir a la (en muchas ocasiones) frustrante y poco rentable estrategia del ensayo y el error.

Notas

[1] La razón que justifica la no inclusión en este trabajo de la pragmática dentro de la gramática responde principalmente a un intento de acotación de los aspectos tratados en el artículo. De todas formas, reconocemos el interés que pueden despertar los aspectos pragmáticos del lenguaje de interrogación. Entre esos aspectos cabe destacar los procesos implicados en la traducción del enunciado del lenguaje natural que representa la necesidad de información a la ecuación de búsqueda del lenguaje de interrogación del SI.

[2] Es importante señalar que a lo largo de este artículo vamos a considerar como términos del LI cualquier palabra o frase de la lengua que ha sido utilizada en los documentos susceptibles de ser recuperados por el SI. Esto supone que el conjunto de términos o frases que se utiliza en los procesos de indización no está limitado, sino que puede contener tantos términos como palabras constituyen un lenguaje natural, lo que significa que una parte del texto de los documentos gestionados se almacena en el ordenador y que la base de datos que se crea a partir de los mismos puede ser interrogada utilizando para ello palabras y frases que aparezcan en el propio texto (5). No hay que olvidar que, alternativamente a la utilización del lenguaje natural, también existe la posibilidad de que se haga uso de los *vocabularios controlados* por parte de los SI (11). Los sistemas que utilizan vocabularios controlados en la gestión documental tienen limitado el conjunto de términos que deben ser utilizados en la indización. De todas formas, nosotros vamos a obviar esa posibilidad en este trabajo por motivos de simplicidad expositiva, ya que si no, nos veríamos obligados a introducir y recordar continuamente cláusulas de restricción que delimitasen el léxico del LI.

Es importante tener en cuenta también que, por un lado, aunque pueda contemplarse otro tipo de operadores para acotar, en este artículo sólo vamos a tratar los booleanos (y, brevemente, los de proximidad) y, por otro lado, que este trabajo está planteado en términos de metacódigo. Esto último significa que aunque en él se utilicen como forma de los operadores booleanos las expresiones «[AND]», «[OR]» y «[NOT]», es posible sustituir éstas por cualquiera de sus formas equivalentes. Así, por ejemplo, «[AND]» puede sustituirse por la expresión «&», «[OR]» por la expresión «|» y «[NOT]» por la expresión «¬».

[3] Definir por *comprensión* un conjunto es ofrecer las propiedades necesarias y suficientes que dan cuenta de todos y cada uno de los elementos que lo conforman. En cambio, definir por *extensión* un conjunto es ofrecer el listado de los elementos que lo constituyen. En este sentido, por ejemplo, puedo definir el conjunto A por extensión diciendo que A = {2, 4, 6, 8} o por comprensión diciendo que A = {x: x es un número par menor que 10}.

[4] Intuitivamente podría decirse que la principal propiedad semántica de una ecuación de búsqueda es la de expresar o definir el conjunto de documentos que es relevante para la necesidad de información cuya correcta traducción se corresponde con esa ecuación de búsqueda y que, por tanto, puede defenderse también que ésta última representa, en cierta medida, una necesidad de información. Esto permitiría aprovechar de un modo más directo todo el edificio teórico y conceptual que se fundamenta alrededor del tema de las necesidades informativas. De todas formas, hemos decidido plantear el componente semántico en términos exclusivamente conjuntistas, sin apelar directamente al tema de las necesidades, por motivos de simplificación expositiva y, sobre todo, para evitar que, de esta manera, puedan deslizarse cuestiones pragmáticas relacionadas con las necesidades de información que se escapen del ámbito exclusivo de la semántica.

[5] Lo mismo ocurre si introducimos un cambio de paréntesis dentro una expresión matemática, como por ejemplo «(2 . 4) + 5», en la que se combinan algunos operadores (la suma y la multiplicación) y obtenemos la expresión «2 . (4 + 5)». De esta manera, mientras que (2 . 4) + 5 es igual a 13, observamos que 2 . (4 + 5) es, sin embargo, igual a 18.

[6] [[((Política [AND] Economía) [OR] Siglo XIX)]] = {d: d ∈ [[(Política [AND] Economía)]] o d ∈ [[Siglo XIX]]}.

[7] [[(Política [AND] (Economía [OR] Siglo XIX))]] = {d: d ∈ [[(Política]] y d ∈ [[(Política [OR] Siglo XIX]]}.

[8] Evidentemente, esta estrategia no garantiza completamente que el ruido y el silencio coincidan con el conjunto vacío, pero sí que ayuda a reducir en gran medida las dimensiones de estos dos conjuntos.

Bibliografía

(1) FRANTS, V. y BRUSH, C. The Need for Information and Some Aspects of Information Retrieval Systems Construction. En *Journal of the American Society for Information Science* (39), 2, 1998, págs. 86-91.

(2) CODINA BONILLA, Ll. *Sistemes d'informació documental: concepció, anàlisi i disseny de sistemes de gestió documental amb microordinadors.* 1993. Editorial Pòrtic S.A. Barcelona.

(3) CODINA BONILLA, Ll. *Tratamiento informatizado de la información: introducción conceptual y tutorial de prácticas,* 1998. Barcelona, Universitat Pompeu Fabra.

(4) RODRÍGUEZ, J.; DÍAZ, P. y PARDO, M. Modelos y estrategias para la recuperación de información por similitud semántica. En *Actes de les 6es. Jornades Catalanes de Documentació de lógica formal.* 1997. Medusa. Barcelona.

(5) LANCASTER, F. W. *Vocabulary Control for Information Retrieval.* 1992. Information Resources Press. Illinois.

(6) MORRIS, C. *Writtings on the General Theory of Signs,* 1971. Mouton, La Haya-París.

(7) GARCÍA-CARPINTERO, M. *Las palabras, las ideas y las cosas. Una presentación de la filosofía del lenguaje.* 1996. Ariel. Barcelona.

(8) BADESA, C.; JANÉ, I. y JANSANA, R. *Elementos de lógica formal.* 1998. Ariel. Barcelona.

(9) BERGMANN, M.; MOOR, J. y NELSON, J. *The Logic Book.* 1990. McGraw-Hill. New York.

(10) QUESADA, D. *La lógica y su filosofía.* 1985. Barcanova. Barcelona.

(11) VAN SLYPE, G. *Les languages d'indexation: conception, construction et utilisation dans les systèmes documentaires.* 1987. Les Editions d'Organisation. París.