>>> revista tradumàtica
Traducció i Technologies de la Informació i la Comunicació
Número 2 : La Documentació : Novembre 2003 : http://www.fti.uab.es/tradumatica/revista :
ISSN 1578-7559

# El lenguaje en la comunicación y recuperación de información[1]

# Language in communication and information retrieval

**Dr. Mario Pérez-Montoro Gutiérrez**
**Estudis de Ciències de la Informació i de la Comunicació, Universitat Oberta de Catalunya**
**Internet Interdisciplinary Institute**

**Resumen**

El lenguaje mantiene una relación estrecha con la construcción de conocimiento. Éste juega un papel crítico en los contextos comunicacionales en los que se transmite información y en las operaciones de recuperación de información. En este artículo se describe, por un lado, el papel y las propiedades gramaticales que permiten que el lenguaje natural intervenga en los contextos de transmisión de información. Y, por otro, se ofrece una descripción paralela del papel y las propiedades gramaticales que permiten que el lenguaje de interrogación intervenga en los contextos de recuperación de información.

**Palabras Clave**: lenguaje, lenguaje de interrogación, conocimiento, comunicación, recuperación de información, sintaxis, semántica.

**Resum**

El llenguatge té una relació estreta amb la construcció de coneixement. El coneixement juga un paper crític en els contextos comunicacionals en què es transmet informació i en les operacions de recuperació d'informació. Aquest article descriu, per una banda, el paper i les propietats gramaticals que permeten que el llenguatge natural intervingui en els contextos de transmissió d'informació. I, per l'altra, ofereix una descripció paral·lela del paper i les propietats gramaticals que permeten que el llenguatge d'interrogació intervingui en els contextos de recuperació d'informació.

**Paraules clau**: llenguatge, llenguatge d'interrogació, coneixement, comunicació, recuperació d'informació, sintaxi, semàntica.

**Abstract**

Language has a very close relationship with the construction of knowledge. It is an essential variable in communication of information and information retrieval. This article offers a description of the role of language in those processes. On the one hand, it describes the grammatical properties that explain the role of natural language in the context of information transmission. And, on the other, it shows the grammatical properties that explain the role of query language in the context of information retrieval.

**Key Words**: Language, Query Language, Knowledge, Communication, Information Retrieval, Syntax, Semantics.

## 1. Introducción

Uno de los rasgos característicos que definen al ser humano, y que lo diferencian de la mayoría del resto de las especies que pueblan el planeta, es su capacidad de conocer y aprovechar ese conocimiento en su propio beneficio.

Sin embargo, el conocimiento obtenido como fruto de esa capacidad no se construye de forma espontánea en el vacío, sino que se fundamenta, en gran medida, sobre la información que llega a nuestro sistema nervioso procedente del entorno que nos rodea. En este sentido podemos decir que la información juega un papel muy importante en la creación del conocimiento.

La información suele verse involucrada de dos maneras distintas en los procesos de gestación de conocimiento: a partir de su transmisión directa y, más recientemente, mediante su recuperación. Los seres humanos solemos construir nuestro background de conocimiento a partir de los intercambios (transmisión) de información que establecemos en la comunicación con otras personas. Pero también, y de una manera más intensiva desde hace unas décadas, aumentamos nuestros conocimientos a partir de la recuperación de información que se encuentra previamente almacenada en repositorios especializados.

Dentro del contexto de la creación del conocimiento, es importante señalar que el lenguaje es una variable fundamental de todo este proceso. El lenguaje juega un papel crítico, pues habilita los contextos comunicacionales en los que se transmite información y estructura (en forma de lenguaje de interrogación) las operaciones de recuperación de información.

En este artículo se intenta describir la doble responsabilidad del lenguaje en el contexto de la construcción del conocimiento. Por un lado, en el apartado número dos, se presentan las propiedades gramaticales que en cierta manera permiten que el lenguaje natural intervenga de un modo decisivo en los contextos de transmisión de información. Y, por otro, en el apartado número tres, se ofrece una descripción paralela del papel y de las propiedades gramaticales que permiten, al menos en parte, que el lenguaje de interrogación intervenga en los contextos de recuperación de información.

## 2. El lenguaje en la comunicación de información

Dentro de los contextos comunicacionales donde se produce transmisión de información, el lenguaje (natural) debe ser considerado como una de las variables críticas. Si se restringiera el uso del lenguaje, nuestra capacidad comunicativa se vería sensiblemente reducida. Y es que los seres

humanos articulamos y vehiculamos la mayoría de los episodios de transmisión de información mediante el uso de una lengua. Cuando nos comunicamos cara a cara con otra persona o cuando lo hacemos mediante el teléfono, el correo electrónico o una simple carta, siempre utilizamos el lenguaje para hacer posible esa comunicación.

Sin embargo, aunque el lenguaje esté presente en los actos comunicativos más cotidianos y cercanos, esto no significa que el papel que juega en este tipo de contextos pueda ser considerado menor. Para entender adecuadamente el peso del lenguaje en los procesos de comunicación, sólo tenemos que revisar los elementos que intervienen en este tipo de procesos.2

En términos generales, podemos decir que en todo contexto comunicativo en el que se produce un flujo o una transmisión de información pueden identificarse, por un lado, una serie de elementos (el emisor, el mensaje$_1$, el contenido informativo, el transmisor, la señal emitida, el canal de comunicación, la señal recibida, el transmisor inverso, el mensaje$_2$, el receptor, la situación, el contexto, el código y el feedback) dotados de unas funciones determinadas y, por otro lado, un factor de disfunción (la fuente de ruido).

En concreto, este proceso puede describirse de la siguiente manera. El emisor (la fuente de información, el elemento del que parte el proceso comunicativo, y que se caracteriza por su intención de transmitir cierta información a un receptor), selecciona un mensaje concreto o una cadena de mensajes (una información determinada, en definitiva, a la cual nos referiremos utilizando la expresión “mensaje$_1$”) de entre un conjunto de mensajes posibles (o de cadenas de mensajes posibles).

En el caso de la comunicación humana, el emisor actúa a su vez de transmisor encargándose de transformar o traducir, a partir del conocimiento de una lengua, ese mensaje produciendo para ello una palabra o un enunciado (señal emitida) que a la postre es la que será transmitida. Esta señal emitida debe transportar la información que coincide con el mensaje$_1$. Para que esa operación tenga éxito, el producto de la traducción, la señal emitida, debe ofrecerse de acuerdo con un código y en sintonía con cierto contexto.

Un código no es más que un sistema de señales gobernadas por un conjunto de reglas que determinan cómo y en qué contextos las señales pueden ser usadas y combinadas para transmitir información. Cualquiera de las lenguas naturales que hablamos (catalán, castellano, inglés, etc.) con sus reglas sintácticas y semánticas puede ser considerada como ejemplo de código. El contexto, en cambio, debe identificarse como el conjunto de señales que ya han sido trasmitidas anteriormente y que preceden a la señal emitida.

La señal emitida circula por el canal de comunicación, adecuándose al mismo y partiendo desde el transmisor. El canal no es más que el medio material cuya alteración permite esa transmisión de la señal emitida. El aire (cuando hablamos en persona con alguien), el papel (cuando escribimos a alguien) o el cable telefónico (cuando conversamos utilizando un teléfono) son ejemplos de canal de comunicación.

La señal emitida llega a través del canal al transmisor inverso. Cuando alcanza esa posición pasamos a denominarla señal recibida. Esta señal recibida acostumbra a coincidir con la señal emitida. El transmisor inverso (una especie de transmisor con la función comunicativa invertida) se encarga de descodificar, transformar o traducir esa señal recibida y convertirla en el mensaje$_2$. Recordemos que este mensaje$_2$, si no se ha producido nada que altere el proceso, coincide con la información que transporta la señal recibida, con la transportada por la emitida y, por tanto, con el mensaje$_1$. Para que esa operación tenga éxito, el producto de la traducción, la señal emitida, debe ofrecerse de acuerdo al código (la lengua utilizada) y en sintonía con el contexto. El receptor, destinatario último del proceso comunicativo y que en el caso de la comunicación humana coincide con el transmisor inverso, recibe e interpreta esa señal final, experimentando cierto efecto a partir de esa información surgida originalmente del emisor y que se ha mantenido a lo largo de toda la cadena de elementos.

Además, en todo contexto comunicativo es posible identificar también otros elementos: la situación y el *feedback*. La situación debe entenderse como todo el entorno espacio-temporal en el que se produce la comunicación. El conocimiento de la misma también contribuye a la correcta actuación comunicativa del emisor y a la correcta interpretación por parte del receptor. El feedback o la retroalimentación debe identificarse como la respuesta del receptor, dirigida hacia el emisor, respecto al efecto que le produce la información que recibe.

Por último, hemos de señalar que, desgraciadamente, en muchas ocasiones, el proceso global de la comunicación que acabamos de describir no se produce tal y como esperábamos. En algunas situaciones concretas interviene una fuente de interferencias, distorsión o ruido que produce una diferencia significativa entre la señal original emitida por el transmisor y la que finalmente recibe el receptor. Esta diferencia entre las señales provoca que el contenido informativo que llega al receptor no coincida con el que originariamente fue seleccionado en la fuente de información o emisor, con lo que se frustra el proceso o el intento comunicativo.

Ahora, una vez que hemos introducido los elementos que conforman un contexto comunicativo, el papel crítico del lenguaje en todo el proceso parece claro. Para que se produzca realmente comunicación, el transmisor (o emisor, ya que es éste el que desarrolla esta función) debe representar adecuadamente la información utilizando un código (una lengua), una situación y un

contexto (para acabar de dotar de sentido a la representación) que, a su vez, sean compartidos también por el transmisor inverso (o receptor, ya que es éste el que realiza esta tarea). El incumplimiento de este requisito puede provocar que la información original (o mensaje$_1$) no se corresponda con el mensaje$_2$, no alcanzándose de esta manera el objetivo de este tipo especial de comunicación: transmitir cierto contenido informativo desde el emisor al receptor.

Sin embargo, el papel del lenguaje en la comunicación puede explicarse también en términos gramaticales. Y es que, dejando al un lado el tema de la pragmática, podemos afirmar que existen dos tipos de propiedades gramaticales que permiten que el lenguaje pueda utilizarse en contextos comunicacionales donde se produce transmisión de información: alguna de sus propiedades sintácticas (derivadas de las relaciones que se producen entre los signos de ese lenguaje) y ciertas propiedades semánticas (derivadas de las relaciones que mantienen esos signos con los objetos que representan).

Comencemos describiendo esas propiedades sintácticas. Los lenguajes (naturales) se encuentran formados por un conjunto de símbolos que recibe el nombre de "léxico de la lengua". 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 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 (García-Carpintero, 1996). 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 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 sintácticas 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 se aplica sobre un conjunto infinito de entidades.

### 3. El lenguaje de interrogación en la recuperación de información

La evolución cultural y tecnológica que hemos protagonizado en la segunda mitad del siglo pasado, no solo nos ha permitido construir conocimiento a partir de la transmisión de información que se produce como fruto de la comunicación directa entre las personas, sino que también ha habilitado la creación de nuevo conocimiento a partir de otra operación cercana pero no coincidente: la recuperación de información.

Estrictamente hablando, la recuperación de información no puede ser identificada como un tipo especial de contexto comunicativo. Los procesos de recuperación no están protagonizados directamente por personas, sino que se establecen entre una persona y un recurso tecnológico (un sistema de información). Sin embargo, dentro de estos procesos el lenguaje (un lenguaje especial: el lenguaje de interrogación) también continúa teniendo un papel crítico.

Desde hace aproximadamente una década, sobre todo desde la implantación de internet y la popularización de los sistemas de gestión de bases de datos, las operaciones de recuperación de información son cada día más habituales y cotidianas. 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.

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 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 compara3 la ecuación de búsqueda con las representaciones de los documentos —obtenidas a partir de un proceso de indización (Lancaster, 1992)—, 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 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 el apartado anterior pudimos describir los dos tipos de propiedades gramaticales que permiten que el lenguaje natural pueda utilizarse en contextos comunicacionales donde se produce transmisión de información. En la misma línea, pero ahora en el escenario de nuestros intercambios con los SI, cabe preguntarse cuáles son las propiedades gramaticales que permiten (al menos en parte) que un lenguaje de interrogación intervenga adecuadamente en una operación de recuperación de información. El aprovechamiento y la adaptación de las ideas del apartado anterior al contexto del lenguaje de interrogación nos pueden dar la respuesta.

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 de términos y por los operadores booleanos [AND], [OR] y [NOT].4 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.

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 se aplica 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 comprensión5 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*.

Por último, nos queda señalar que, 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 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 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.

**4. Bibliografía**

BADESA, Calixto, Jané, Ignacio y JANSANA, Ramon (1998). *Elementos de lógica formal*. Barcelona: Ariel.

CODINA BONILLA, Lluís (1996). *El llibre digital: una exploració sobre la informació electrònica i el futur de l'edició*. Barcelona: Generalitat de Catalunya, Centre d'Investigació de la Comunicació.

FRANTS, V. y BRUSH, C. (1988). "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.

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

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

POLLIT, A.S. (1989). *Information Storage and Retrieval Systems. Origin, Development and Applications*. Chichester: Ellis Horwood Limited.

SHANNON, C. y WEAVER, W. (1949). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

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

---

[1] Los resultados que aparecen en este trabajo se integran dentro del proyecto de investigación KAIMI (*Knowledge Assets Identification and Methodology of Implementation in Organizational Knowledge Management*) que se está desarrollando gracias a la financiación económica del IN3 (*Internet Interdisciplinary Institute*) de la Universitat Oberta de Catalunya.

[2] Para introducir todos estos elementos vamos a recuperar algunas de las ideas defendidas por Shannon y Weaver en su famosa *Teoría Matemática de la Comunicación*. Teniendo como objetivo la claridad expositiva, nos hemos tomado la libertad de realizar pequeñas adaptaciones y modificaciones sobre esas ideas y completarlas con una serie de elementos (la situación, el contexto, el código y la noción de *feedback*) que no son contemplados por estos autores.

[3] Para introducirse en los principales estrategias —la conjuntista y la vectorial— mediante las cuales los SI comparan las ecuaciones de búsqueda con las representaciones de los documentos puede consultarse Codina,

1998. De forma general, para introducirse en los modelos y estrategias basados en la similitud semántica utilizados en la recuperación de información puede consultarse Rodríguez, Díaz y Pardo, 1997.

[4] Es importante señalar que, por un lado, aunque puedan contemplarse otro tipo de operadores, para acotar, en este artículo sólo vamos a tratar los booleanos 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 estas por cualquiera de sus formas equivalentes.

[5] 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}.