IMPLICACIONES DOCUMENTALES EN EL PROCESAMIENTO DEL LENGUAJE NATURAL

José A. Moreiro González*.

*Publicado originalmente en: CIENCIAS DE LA INFORMACIóN, 24(1) marzo 1993

Los comienzos de una relación íntima

Voy a empezar con una evidencia: el sometimiento a procesamiento automático del lenguaje implica fundamentalmente la presencia del ordenador. Por lo que sólo desde la irrupción informática podemos hablar de dispositivos (ordenador, programas, soportes magnéticos de textos, escáner de lectura, ...) que cambiaron radicalmente el manejo de los productos lingüísticos y que a la larga nos concedieron lo que hoy engloban las **industrias de la lengua.** Dentro de ese marco y en un paralelismo temporal y metodológico se sucedieron tanto la revisión de las teorías lingüísticas, como un nuevo modo revolucionario que modificó radicalmente la manera del hacer documental. Se abría la era de la *" Information Science"*, que cambio de forma profunda las teorías y técnicas de nuestro ámbito (1).

El paralelismo se inicio, pues, con la llegada de los ordenadores y se ha mantenido hasta nuestros días dentro de cauces muy próximos. Si hacemos un recorrido somero, veremos cómo hay ya una primera coincidencia desde el inicio de la automatización y la subsiguiente revisión de las teorías lingüísticas, que afectó tanto a las aplicaciones propiamente lingüísticas, como a las documentales.

Desde los años 60 se ensayaron aplicaciones de traducción automatizada en total ajuste con la utilización de los ordenadores en los procesos documentales, causa del desarrollo acelerado que nuestro campo ha vivido tanto en la investigación y en la profesión, como en la formación y en la empresa (2). Tuvo como consecuencia la aparición de unos hechos documentales que marcaron una situación absolutamente nueva:

- En primer lugar, nuevos procesos de tratamiento (indización por descriptores y lenguajes coordinados), que implicaron que la problemática lingüística entrase de lleno en la gestión de los documentos (3).
- Luego, una nueva filosofía de actuación (4): manejo más rápido y exacto de la información. Posibilidad de difusiones más pertinentes. Mediatización del tratamiento documental por el uso de las tecnologías. Aparición de las bases de datos.
- Además, el ordenador no sólo manejaba descriptores o hacía tarea de ordenación, sino que, además, podía realizar trabajo intelectual referente al análisis de contenido de los textos escritos. Fue la propuesta de Luhn sobre indización automática desde métodos relativos a la característica de frecuencia y a la localización de palabras en los textos (5). El propio Lunh desarrollaba a la par técnicas automáticas de resumen e índices tipo KWIC (Key word in context).
- También tuvo relación directa con el sometimiento de los hechos informativos a medidas estadísticas, conformándose como el principal apoyo de los métodos bibliométricos (6).

¿En qué contexto advino esta aproximación de la lengua a la gestión de documentos?

1) Causado ante todo por las exigencias sociales de comunicación rápida y efectiva, que requerían procesos

que acelerasen la traducción de los documentos (de una lengua a otra, de un texto a sus representaciones).

2) Por la apertura de unos países a otros, conforme se habían ido superando las consecuencias de la II Guerra Mundial y se accedía a un intercambio constante de personas, ideas y mercancías.

Si en los años 60 comenzaba la documentación a acercarse a la lingüística, no fue hasta la década de los 80 cuando la inteligencia artificial hizo que la zona de contacto fuese mayor.

En el intermedio se fue precisando que las estructuras de la lengua fueran interpretadas mediante algoritmos lógicos. Y aquí vuelven a coincidir los procesamientos lingüísticos para versar la distribución, las estructuras y los términos de una a otra lengua con los paralelos de una expresión lingüística, lo que entendemos por un texto o documento, hacia su representación conceptual o textual reducida (7). El desarrollo de la lingüística del texto se impulsó desde las dificultades de versión surgidas de la diferente distribución de los elementos de una lengua a otra. Aumentadas por el «desorden» de presentación gramatical, por la posibilidad de connotaciones y ambigüedades en las palabras, y por las diversas funciones que una misma palabra puede cumplir; así como derivadas de la comprensión del contexto y de la situación de cada acto lingüístico, de las anáforas (relación antecedente-consecuente), e, incluso, de las peculiaridades personales de expresión, todas ellas dificultades idénticas para uno u otro objetivo

La llegada de la inteligencia artificial
en los años 80

Los años 80 trajeron los primeros programas informáticos que permitían alcanzar soluciones para los problemas antes relacionados. La posibilidad de aproximación al procesamiento deseable y / o aceptable quedaba abierto. La irrupción de la inteligencia artificial en esa década aproximó aún más la informática a la lingüística. Los ordenadores empezaban a simular el comportamiento cognitivo, posibilitando su presentación lógica y lingüística. Los espacios mentales en aplicación mecanizada potenciaron la práctica automatizada de los fenómenos lingüísticos. Se utilizaron inmediatamente en :

- la traducción,
- los trabajos del lenguaje,
- el análisis y la generación de textos y, causados por la capacidad de la máquina para responder al lenguaje natural,
- los analizadores y sintetizadores de voz y escritura.

Para ello, los ordenadores tenían que conocer, al menos:

- las estructuras gramaticales,
- las sinonimias,
- la semántica de mundo (puesto que se trabaja a partir de inferencias), y
- las estructuras textuales.

En resumen, tanto para traducir como para tratar documentalmente los textos se ha requerido desde entonces conocer previamente las bases lingüísticas de éstos.

Pese a todo, una década después vemos que la intervención correctora humana sigue siendo necesaria, y que los tipos de documentos que se someten a prueba conforman un porcentaje muy pequeño. Sería el caso de los documentos técnicos de superestructuras normalizadas y terminología muy estable; o también muchos documentos administrativo-legales (por ejemplo, los de la Comunidad Económica Europea [CEE]). Por otra parte, las investigaciones se han tenido por rentables sólo cuando se han aplicado a un número de documentos muy elevado.

el sometimiento a procesamiento automático del lenguaje implica fundamentalmente la presencia del ordenador.

Los sistemas experimentados hasta el momento procesan los textos en campos de aplicación muy limitados, y desde tipos textuales muy modelizables, cuyas características son predecibles y bien comprendidas. Por lo que la lista de aproximaciones es aún muy restringida. El éxito se complica mucho cuando consideramos la posibilidad de tratar materiales muy diversos de todos los campos del conocimiento.

Implicaciones documentales del procesamiento del lenguaje natural

La palabra es el signo abstracto por cuyo medio se

expresan los hombres. Los problemas para trasladar y controlar las representaciones semánticas parten de la palabra, pero, asimismo, en ella encuentran solución. Ahora estamos en el momento en que las palabras pueden leerse, registrarse, comprenderse y representarse en el ordenador.

Desde una perspectiva interdisciplinaria (Inteligencia artificial, socio y psico-lingüística, filosofía del lenguaje, lógica, filología, ciencia del conocimiento) se ha hecho de la lengua una materia trabajada por la industria y el comercio. La automatización de sus procesos presenta claras consecuencias para la documentación. Ya que no en vano ésta, por su origen, características y destino es antes que otra cosa un hecho lingüístico: El análisis documental y sus intentos de automatización se sitúan de lleno en una panorámica lingüística caracterizada por su entronque con los lenguajes científico-técnicos de las áreas del conocimiento. Si el fin de la documentación es posibilitar principalmente las comunicaciones científicas, el de las Industrias de la Lengua es posibilitar el intercambio y la cooperación científicas sobre todo. La coincidencia de fines es grande. No debe extrañarnos: los documentos se expresan mediante las lenguas. El documentalista trata lenguas naturales desde el análisis de los textos en que se reflejan los hechos de la vida. Para analizar y manejar esos documentos, poderles indizar o generar nuevos textos desde ellos, deben previamente representarse de manera sistemática las lenguas naturales (8).

Incluso, ha sido paralelo el desarrollo de técnicas aplicadas a la traducción automatizada y al resumen mecánico. Cuando los ordenadores de la primera generación traducían palabra a palabra y se preocupaban por las equivalencias sintácticas (SYSTRAN), los extractos se confeccionaban con criterios terminológicos. La segunda generación de ordenadores se preocupó de los factores semánticos tanto en la extracción como en la traducción. La inteligencia artificial se ha centrado en comprender los textos y tiende a aplicaciones de factores pragmáticos en la traducción (9). Hemos razonado antes que los problemas en la versión de una lengua a otra son los mismos que se dan en el tratamiento de contenido documental. Y si en marco de la automatización de las traducciones bajo tutela de la CEE nació el concepto de *industrias de la lengua* , obligatoriamente deben incluirse en éstas los procesos lingüístico-documentales, pues tienen que ver directamente con el tratamiento automatizado de las lenguas naturales.

La mayor relación entre la traducción y la documentación se ha dado a lo largo del procesamiento técnico de ambas, que ineludiblemente requiere:

- comprender texto,
- lograr un sistema macroestructural,
- generar un nuevo texto o representación.

Esta vía ha supuesto, además, que la representación semántica tienda a fusionar los procesos de indización y resumen facilitando su logro.

Los sistemas de reconocimiento textual fundamentales para las nuevas tendencias de la recuperación, los apoyos automáticos a la extracción de conceptos y de oraciones, y los intentos de síntesis textuales han obligado al documentalista a participar en el desarrollo teórico y práctico de la lingüística del texto. La comprensión de los textos debe hacerse en contexto y desde aplicaciones inferenciales jerarquizadas. Para lo que se consideran tanto la finalidad de los documentos y el mundo que reflejan, como los objetivos que persigue el documentalista.

> Los algoritmos de traducción (diccionarios terminológicos, más gramática, más logicales) han sido los mismos usados en ambas versiones macroproposicionales de los textos.

Desde luego la gestión documental y la bibliográfica se han aprovechado de los mecanismos automáticos de ayuda a la traducción. Sin embargo, de cuantas aplicaciones directas pueden interesar al manejo de la información nos conviene especialmente cuanto se refiere a las terminologías y al establecimiento de los resúmenes.

Los métodos informáticos posibilitaron la confección automática de las terminologías. En su constitución se hizo común el trabajo en equipo de especialistas en un área, terminólogos e informáticos. Nedobity (10) considera el establecimiento de las terminologías necesario y fundamental en estas actividades:

- ordenar los conceptos científicos;
- formular informaciones específicas;
- transferir el conocimiento en la enseñanza;
- transmitir el conocimiento entre lenguas;
- recuperar e indizar la información.

Así, pues, las estructuras de la lengua natural, mediante la actuación de los ordenadores y de los programas de inteligencia artificial, van a poder aplicarse al análisis y recuperación de los documentos, facilitando la formación, consulta y gestión de las bases de datos.

Los algoritmos de traducción (diccionarios terminológicos, más gramática, más logicales) han sido los mismos usados en ambas versiones macroproposicionales de los textos. Incluso se han fusionado los dos objetivos cuando se produjeron programas, como el TITUS, destinados a la traducción de resúmenes entre varias lenguas.

La necesidad de compatibilizar los numerosos extremos estructurales y estratégicos, que se dan en el texto, nos permiten subrayar estos factores a tener en cuenta en un modelo de análisis a automatizar (por más que deban verse globalmente, no parcelados como en esta numeración proponemos):

- Morfología: tanto flexiva como derivacional.
- Lexicografía: ortografía de los morfemas, aspectos morfológicos, sintácticos y semánticos de cada morfema (precisión de la semántica de mundo). Problemas de polisemia (homonimia) y de paráfrasis (sinónimos y definiciones). Mediante la lexicografía se posibilita el establecimiento en cada materia de diccionarios y bases de datos terminológicas especializadas. En la traducción realizan la correspondencia de palabras de una a otra lengua. En documentación son la base para establecer los *Thesauri* , verdadero lenguaje de interpretación indical.
- Modelos algorítmicos de sintaxis: fomalizaciones gramaticales .
- Redes semánticas: representación de los co nocimientos.
- Interpretación (inferencias) :

• Desde el contexto: Los marcos. Las anáforas lo no dicho. Todo ello hacia el sentido completo y la eliminación de ambigüedades. Aclaración de los tropos.

• Desde la referencia personal: La función personal, su objetivo, el nivel de competencia (la información previa dominada por los comunicantes). El contorno comunicativo y la situación.

- Las aplicaciones pragmáticas: lo retórico y lo convencional en cada lengua.

La solución de los problemas se efectúa desde un modelo de situación, no directamente desde un texto. Al modelo se aplican las matemáticas y la lógica, y al haberse confeccionado como resultado de todo el proceso de interpretación actúa como base de ulteriores operaciones cognitivas (11). Esta acumulación de experiencias permitirá que los programas aprendan, lo que de momento sólo una aspiración y un límite más la tratamiento automatizado.

Productos documentales e industrias de la lengua

Junto a los diccionarios terminológicos podemos aportar una amplia nómina de productos surgidos en la actividad industrial de la lengua y que tienen aplicación inmediata en documentación:

- *Indización automática* : preparación mecánica de los documentos para la recuperación desde combinacio
nes algorítmicas y de comparación con diccionarios de términos (tanto vacíos como relevantes).

Incluso se han fusionado los dos objetivos cuando se produjeron programas, como el TITUS, destinados a la traducción de resúmenes entre varias lenguas.

- *Aplicaciones a la interrogación de bases de datos* : mediante interfaces entre lenguajes informáticos y lengua natural.

- *Tratamiento y almacenamiento de grupos de textos* : hacia el documento global y las relaciones temáticas de documentos independientes (hipertexto [12], y teorías de Kochen [13]).

- *Conocimiento de las estructuras textuales* : imprescindibles para formar extractos, dotar de profundidad y relevancia a la indización, posibilitar el resumen automatizado, e incluso explicar un método consistente de resumen humano (14).

- *Apoyo a la investigación en los aspectos formales* : función lógica de la documentación, que se extendería hasta las correcciones ortográficas y de presentación de documentos.

Marco lingüístico para una teoría de la documentación

La lingüística no sólo afecta a los aspectos teórico-aplicativos de la representación documental, ya que hasta los propios conceptos fundamentales de nuestro campo encuentran acogida en el cuerpo de principios que conforman su área de conocimientos. Desde ella planteamos unas guías reflexivas:

No cabe duda de que la orientación del *information science* supuso para la documentación un cambio de rumbo. La irrupción de los ordenadores llevó consecuentemente al olvido de la traducción erudita, prefiriéndose planteamientos más tecnicistas. La documentación paso a contemplarse como un sector productivo, y sus métodos se dispusieron en este sentido. Consideremos cómo los primeros teóricos de la *Information science* provenían del mundo informático, y desde una concepción matemática (15) corrobora cómo la explicación de nuestro campo, desde los presupuestos más fundamentales, partía del mundo de la ingeniería de las comunicaciones. Si a ello añadimos el papel fundamental de los ordenadores y de las empresas que potencian el mercado documental, comprenderemos por qué aún hoy la mayoría de las contribuciones en las revistas del sector son bibliométricas, y los modelos buscados están más cerca de lo técnico que de las humanidades.

Sin embargo, no es infrecuente el planteamiento de organismos teóricos provenientes del mundo de las humanidades. Dentro de las cuales los cuerpos de mayor coherencia provienen de explicaciones nacidas de la lingüística.

Creo que desde sus fundamentos no sólo se explican los fenómenos del lenguaje implicados en la comunicación documental (que por evidentes se defienden sin mucha dificultad), si no que se alcanza el razonamiento final sobre qué sea la documentación, sus fines y medios.

Se supera de esta forma el viejo dilema que discute si nuestro campo es una ciencia o una técnica, y a la par se logra un cuerpo de doctrina coherente y sólido:

-Desde esta perspectiva podemos explicar qué lugar ocupa la documentación en el Discurso general de la ciencia y en el texto concreto, igual que en el resto de las comunicaciones humanas en que interviene (16). Se aprecia el valor pragmático de la documentación en el logro de la comunicación de las informaciones originadas por el hombre, en especial dentro del conocimiento científico.

-Se dota de argumentos la presencia necesaria de la automatización en los procesos, y la intervención de las tecnologías y sus productos como apoyo a la elaboración técnica de los mismos. El ordenador es una máquina de información, que mediante operaciones lógicas transforma las entradas de datos en salidas de datos, por lo que representa unos productos de la facultad semiótica humana para construir sistemas de signos (17).

-Se afirman doctrinalmente las bases para enfrentarnos al principal problema de la documentación: la intermediación informativa y la consiguiente versión representativa de los documentos. A la vez que se aclaran las barreras que impiden la recepción de los mensajes existentes (18).

-Se impulsan las perspectivas psicológicas y sociológicas que intervienen en un proceso comunicativo-social como el nuestro (19).

-Se afianzan las tendencias hacia la globalidad de aplicaciones y hacia la globalización de los resultados (informes sintéticos, documentos globales sobre una situación o hecho, visiones generales de un problema o situación científica o humana (20),

-Incluso se entienden los enlaces discursivos entre el centro o empresa documental y los usuarios, sus necesidades, y la oferta y difusión de los productos.

Creo, por tanto, en la validez de un modelo lingüístico para explicarno sólo los fenómenos lingüísticos de la documentación, sino la existencia pragmática de ésta, y por supuesto sus fundamentos conceptuales. Ya que hasta el enfoque interdisciplinario de nuestro campo encuentra su nexo principal precisamente en la lengua, cuyo estudio es, interdisciplinario.

Conclusiones

El lenguaje natural se aprecia como referencia constante para un documentalista. El teórico de la indización se preocupa de establecer relaciones entre el lenguaje natural y los lenguajes documentales, estos se originan en aquel por más que luego hayan constituido una sintaxis particular. El resumen, por su parte, al pasar del original a su explicación reducida se mueve siempre dentro del lenguaje natural.

El objeto al que atiende la documentación es semiótico, y se hace fehaciente mediante el lenguaje. Partimos de discursos, y llegamos a productos documentales cuya estructuración, métodos de análisis y representación se realizan por medios lingüísticos.

La actividad analítica documental demuestra un modo de competencia lingüística. A través de una estructura y siguiendo unos objetivos determinados representamos un discurso con especiales características semántico-pragmáticas. Que además surgen como consecuencia de los propósitos comunicativos de la ciencia y de las intenciones y motivaciones científicas del documentalista, que busca una interacción con los posibles usuarios.

Dado el carácter analítico-sintético de la descripción documental, debemos destacar que los procesos inferenciales, las representaciones, las expectativas y los conceptos memorísticos influyen tanto en la creación como en la comprensión textual. El desarrollo constructivo textual se somete a regulaciones de carácter lingüístico y retórico, así como a operaciones de abstracción y sumarización. Por lo que tiene gran importancia considerar la intervención estratégica del conocimiento. Las relaciones lógicas y psicológicas, y los modelos cognitivos permiten tanto comprender como producir un documento.

Referencias

1) Moreiro González, J. A. Introducción bibliográfica y conceptual al estudio evolutivo de la documentación. Barcelona: PPU, 1990. pp. 169-190.

2) Warner, J. Semiotics, Information Science, Documents and Computers, Journal of Documentation, 46 (1): 16-32, 1990.

3) Kobachi, N. I. Análise documentaria. Consideraçiões sobre un modelo lógico-semántico. En GRUPOTEMMA. Análise documentaria. Consideraçiões teóricas e experimentaçiões. Sao Paulo: FEBAB, 1989. p.47.

4) Vickery, B. y Vickery, A. Information Science in Theory and Practice. Londres: Butterworths, 1987. pp. 132-140.

5) Luhn, H. P. De automatic creation of literature abstracts, IBM Journal of research and development, 2 (2): 159-165, (1958).

6) Terrada, M. L. y López Piñero, J. M. Historia del concepto de documentación, documentación de las ciencias de la información, IV. p. 266. (p. 229-248), (1980).

7) Sagredo, F. e Izquierdo, J. M. Concepción lógico lingüística de la documentación. Madrid: IBERCOM, 1983.

8) Izquierdo Arroyo, J. M. La ciencia de la búsqueda documental secundario a, documentación ciencias de la información, (13): 96-106, (1990).

9) Vidal Benyto, J. La industrialización de las lenguas. En su (dir.) Las industrias de la lengua. Madrid: Fundación G. Sánchez Ruipérez, 1991. p. 14.

10) Nedobity, W. The relevance of terminologies for automatic abstracting. Journal of information Sciencie, (1982), p. 163.

11) Dijk, T. A. Van, y Kintsch, W. Strategies of discurse comprehension. New York: Academic Press, 1983. p. 340.

12) Caridad, M. y Moscoso, P. Los sistemas de hipertexto e hipermedios. Una nueva aplicación en informática documental. Madrid: Fundación Germán Sánchez Ruipérez, 1991. p. 124.

13) Suanson, D. Integrative mechanisms in degrowth of knowledge: A legacy of Manfred Kocochel, Information Processing ang Managemen, 26 (1): 16, (1990).

14) Bernárdez, E. Introducción a la lingüística del texto. Madrid: Espasa-Calpe, 1982.

15) Shannon, C. E. y Weaver, W. The Mathematical theory of comunnication. - Urbana: University of Illinois Press, 1949. y Belkin, N. J. Information Concepts for information science, Journal of documentation, 34 (1): 55-85, 1979.

16) Moreiro González, J. A. Consideraciones acerca de la intermediazión discursiva de la documentación en la ciencia, Bilduma (rentería-Guipúzcoa), (6):147-154, (1992).

17) Warner, J. Semiotics, Information Scienc, documents and computers, Journal of documentation, 46 (1): 16-32, (1990).

18) García Gutierrez, A. Connotaciones lingüíticas para una teoría de la documentación, Revista Brasileira de biblioteconomía e documentaço, 21 (1-2): 10, (1988).

19) Mckenzie, D. F. Bibliographic and sociologic of texts. London: British Library, 1986.

20) ELLIS, D. Theory and explanation information retrieval research, Journal of Information science, 8 (1): 25-38, (1989).