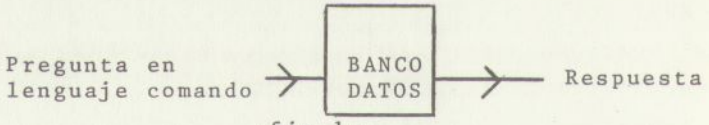
UN SISTEMA PREGUNTA-RESPUESTA EN CASTELLANO, SOBRE UN CORPUS LITERARIO.

Por E. García Camarero y M. F. Verdejo.

1.- Introducción al problema

La organización de la información sobre objetos de un campo específico, con vistas a realizar consultas posteriormente, es el objetivo de algunos sistemas informáticos actuales.

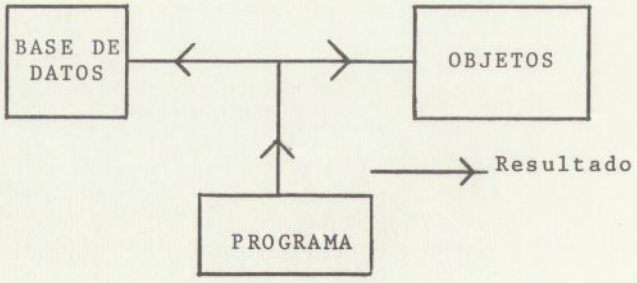
Es el caso, por ejemplo, de una compañia que desea realizar la gestión de la cartera de pedidos de sus clientes. Uno de los procedimientos más utilizados es la construcción de un banco de datos, al que se puede interrogar en un lenguaje de comando, para obtener la información que se encuentra explicitamente almacenada, o que se puede extraer por métodos simples de deducción a partir de los datos que forman el banco. Este proceso está representado por el esquema de la figura 1



-fig.1-

En algunas aplicaciones, los objetos tratados pueden estar almacenados en el ordenador. En este caso se puede obtener más información que la contenida en la base de datos, ya que ciertos programas pueden realizar procesos de transformación sobre los objetos para extraer otro tipo de información.

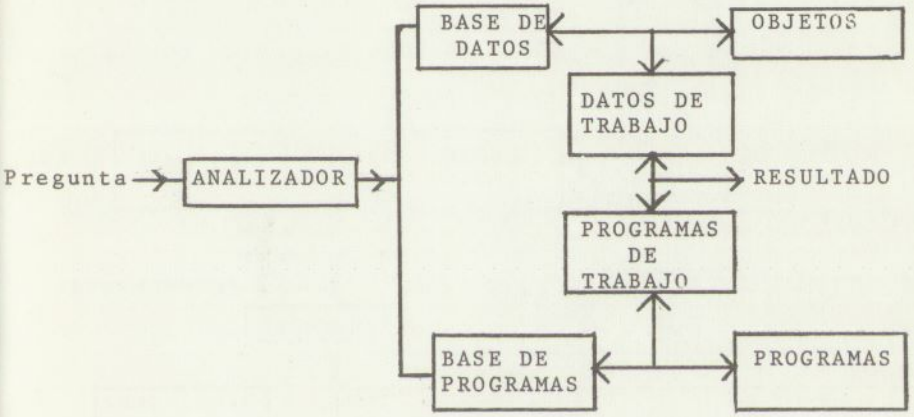
La situación se ilustra por el siguiente esquema:



-fig.2-

Para cada tipo de consulta, será necesario un programa específico que sea capaz de obtener el resultado deseado. Existen al menos dos soluciones: una construir cada vez un programa ad-hoc al problema; otra es la creación de una biblioteca de programas, estableciendo los criterios de selección de cada uno de ellos y definiendo los parámetros necesarios para su ejecución.

Esta operación se puede automátizar de forma que, a partir de la consulta, se pueda determinar el programa y los datos necesarios para producir una respuesta, eliminando así toda intervención humana. El esquema es el siguiente:



-fig.3-

Respecto al lenguaje utilizado para formular las consultas puede escogerse un lenguaje de comando, lo que requiere una cierta formación por parte del utilizador o bien plantearse la posibilidad de utilizar el lenguaje natural, aceptando ciertas limitaciones, a dos niveles: por un lado semánticas (ya que se trata de una aplicación en un campo específico, se supone una comprensión respecto del discurso normal en el dominio); por otro lado sintácticas (para limitar la complejidad gramatical de las frases).

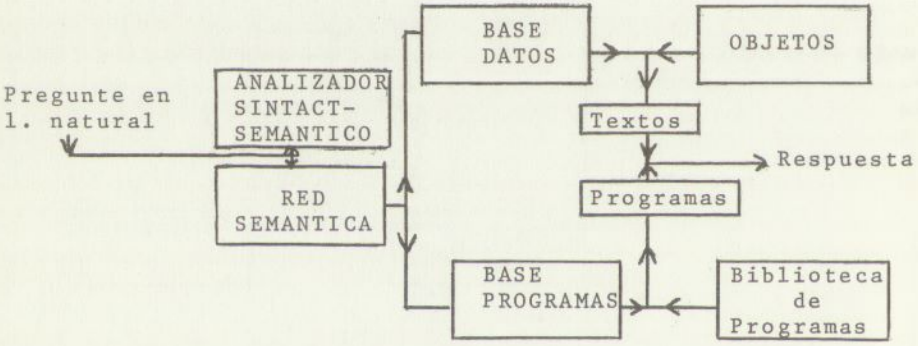
Nuestro objetivo, es desarrollar un sistema de estas caracteristicas para una aplicación específica: El tratamiento de un conjunto de textos literarios para extraer informacion de tipo estadístico.

El subconjunto del castellano que trataremos es el que habitualmente se utiliza para formular preguntas a este respecto.

La primera tarea que el sistema debe resolver es la "comprensión" de las frases en castellano, producidas al formular la consulta. Se trata de determinar qué información se desea y sobre qué objetos. Mediante un proceso de análisis sintáctico y semántico, se extrae el significado de la frase y se representa mediante una notación formal: grafo semántico. Este grafo expresa en forma explicita los objetos implicados así como las acciones (programas) necesarios para producir una respuesta.

La segunda etapa, -resolución del problema- se lleva a cabo por una interacción del grafo, de la base de datos, y de la biblioteca de programas, en un proceso de construcción de la respuesta deseada.

El esquema del proyecto, esta representado por la figura 4:



-fig.4-

El diseño y la implementación de un sistema así concebido exige:

1) Definición de la formalización utilizada para representar y manejar internamente la información: la red semántica.

2) Descripción del proceso de análisis que permita extraer la información de la consulta formulada en castellano y su representación mediante una red semántica.

3) Definición y organización de la base de datos, que constituye la descripción de los objetos.

4) Organización y descripción de los programas.

5) Interacción de datos, programas y red semántica.

2.- Etapas del programa

2.1. La red semántica y el análisis de las frases.

Se puede formular una consulta mediante una frase interrogativa o imperativa.

El análisis de la frase, es el proceso que permite obtener una representación que exprese de forma no ambigua su significado.

La formalización que hemos escogido, es una red semántica en la que los nodos representan objetos, las propiedades que les describen, y las acciones que actuando sobre ello producen ciertas transformaciones.

Por ejemplo, la representación interna de las siguientes consultas:

1.- Listar los contextos de las diez palabras más frecuentes en la poesía de Góngora.

2.- ¿Cuáles son los contextos de las diez palabras más frecuentes en la poesía de Góngora?.

3.- ¿Cuáles son las diez palabras más frecuentes en la poesía de Góngora y en que contexto aparecen?.

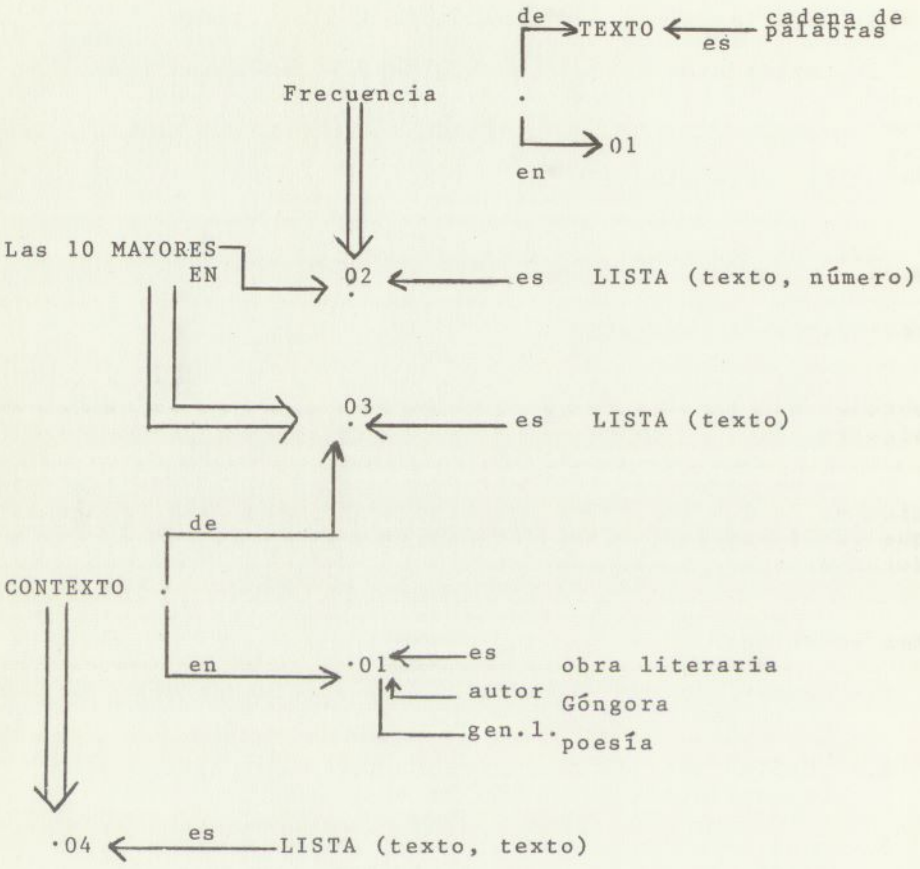
4.- Determinar en la poesía de Góngora las diez palabras más frecuentes y sus contextos.

Es la que damos en la figura 5.

El grafo que representa internamente el significado de las consultas 1,2,2 ó 4, se ha construido a partir de tres subgrafos ACCION.

El primero expresa el cálculo de la frecuencia de las palabras que aparecen en un conjunto de obras escritas por el autor (Luis de Góngora) y el genero literario (poesía).

El segundo describe las diez palabras más frecuentes en el objeto resultado del esquema precedente.



-fig.5-

El tercero expresa el contexto de esas palabras en las obras literarias especificadas.

El proceso de análisis que permite la construcción del grafo, se lleva a cabo, utilizando conocimientos sintácticos, pragmáticos, y un lexico en donde hay información diversa respecto al significado de las palabras: si denotan objeto, atributo, relacciones, ect.

Así por ejemplo, ciertos verbos, como calcular, enumerar, empezar, seguir, continuar y algunos nombres como frecuencia, contexto, longitud, ect. se representan mediante grafos del tipo ACCION, mientras que palabras como letra, párrafo ... ect., que describen conceptos se representan mediante nodos objeto.

Responder a una consulta, a partir del grafo, es determinar el valor de los objetos que vienen descritos mediante atríbutos.

2.2. Base de datos

Para localizar ciertos objetos, se accede a la base de datos, en donde estan descritas las obras literarias que se encuentran en la biblioteca de textos almacenadas en el ordenador.

La organización de la base de datos tiene también forma de grafo semántico, en donde por ej.: una obra concreta, está representada por un nodo, y los arcos, que de él salen o llegan, especifican pares atributo-valor.

2.3. Caracterización de los objetos

Aquellos objetos que se producen como resultado de una determinada acción, pueden describirse formalmente a partir de un conjunto de primitivas. La ejecución imbricada de diferentes programas construye nuevos elementos a partir del conjunto primitivo.

Este, puede describirse, en forma B.N.F. por las reglas siguientes:

letra := a / b / c / .... /z

texto := letra / letra, texto

número := entero / real

lista (texto) := texto / texto, lista (texto)

lista número) := número / número, lista (número)

lista (texto-número) := texto-número / texto-número, lista (texto-número)

lista (texto-texto) := texto / texto, lista (texto-texto)

lista (número-número) := número / número, lista (número-número)

lista := lista (texto) / lista (número) / lista (texto-número) / lista (texto,texto) / lista (número-número)

lista (lista) := lista / lista, lista (lista)

2.4. Organización de los programas

Los programas estan organizados en clases, jerarquizadas según los tipos de entrada que admiten y de salida que producen. A modo de ejemplos, podemos enumerar los siguientes:

FRAGMENTAR (texto) —— lista (texto)

CONCATENAR (lista (texto)) —— texto

AÑADIR (elemento, lista (elemento)) —— lista (elemento)

COMPONER (elemento-elemento)—— lista (elemento-elemento)

DESCOMPONER (lista (elemento-elemento)) —— lista (elemento), lista (elemento)

ORDENAR (lista) —— lista

CONTAR (lista) —— número

LOCALIZAR (lista, elemento) —— número

Cada una de las clases, está formada por un conjunto de programas diferentes, por ej.: para FRAGMENTAR un texto, puede haber distintos criterios:

- Utilizar un separador, en cuyo caso tenemos

  FRAGMENTAR (texto, separador) —— lista (texto)

Si el separador es un blanco, tenemos la fragmentación del texto en palabras. Si es un punto, el resultado es una lista de párrafos, ect.

Otro criterio que puede emplearse es la longitud:

FRAGMENTAR (texto, longitud) —— lista (texto)

Es una operación que produce una lista de textos de longitud definida.

2.5. Interacción del grafo con la base de datos y los programas.

La respuesta se produce mediante la ejecución de las acciones descritas en el subgrafo que representa el significado de la consulta. Así por ejemplo, para el esquema 5, la secuencia consta de los siguientes pasos:

- Determinar (mediante una busqueda) en la base de datos, el objeto descrito por las propiedades:

  Ser obra literaria
  Autor = Góngora
  Genero literario = poesía

  localizarlo en la biblioteca de textos.

- Construir un objeto 02, mediante la ejecución de los programas

  1.- fragmentar un texto en palabras.
  2.- calculo de la frecuencia.

- Construir un objeto 03, lista

  1.- Ordenar crecientemente 02 por el valor de la frecuencia.
  2.- Tomar los 10 primeros elementos.

- Construir el objeto 04

  1.- Comparar un texto con otro.
  2.- Extraer un fragmento de longitud n.

Al final del proceso, tenemos creado el objeto 04, que responde a la consulta formulada.

3.- Conclusión

El sistema, está siendo construido, para un campo específico: la linguística estadística, aunque es posible aplicarlo a la construccion de bases de datos de areas diversas.

Nuestro trabajo está orientado hacia el estudio de una metodología y el desarrollo de una tecnología que permita construir sistemas informáticos integrados que puedan ser empleados por personal no-informático. En esta perspectiva, nos parece fundamental el poder usar el lenguaje natural para expresar las consultas. Lo que no excluye la utilización del sistema mediante un lenguaje de comando.

4.- Bibliografía

CODD, Arelational model of data for large shared Data Banks, CACM 13,6, (June 1970).

DELOBEL, Les systemes de bases de données, Ecole d'été de 1'AFCET 75.

DIJKSTRA, Structured programming, Academic Press 1972.

SCHANK, Conceptual Information processing, North-Holland.

SIMMONS, Semantic Networks: their computation and using for understanding english sentences. En Computers Models of thought and langage, Freeman, 1973 pag. 63-113.

WIRTH, Program development by stepwise refinement, CACM vol. 14, nº 4, Abril 1971.