

I.3 APLICACIÓN DE UN RECONOCEDOR DE LENGUAJE NATURAL RESTRINGIDO A LA RECUPERACIÓN DE DATOS

Gabriel Cordero Sánchez*

Resumen

En este documento se muestra la estructura funcional de un reconocedor de lenguaje natural restringido, como una herramienta (de software) de uso general. También se describe su aplicación en la recuperación de datos (interactiva o reportes), actualmente de subsistemas de información administrativos (Recursos Financieros y Recursos Materiales) desarrollados en la Dirección General de Información Administrativa (DGIA) como apoyo en la administración de la Secretaría de Educación Pública (SEP).

El reconocedor, como herramienta de uso general, tiene la característica de operar a partir de una base de datos, en forma independiente a la aplicación particular que se trate.

La recuperación de datos, se basa en solicitudes de tipo imperativo en lenguaje natural restringido, las cuales pueden ser conocidas o no para el reconocedor debido a que una de sus facilidades es la capacidad de registrar (y en cierta forma aprender) frases a nivel de palabra por palabra, de lo que hasta ese momento son instrucciones de usuario desconocidas.

El reconocedor es un programa interactivo que reconoce frases imperativas mediante validaciones de tipo léxico-sintáctico usando una gramática y como resultado genera directivas de acceso a una base de datos. Es capaz de incrementar su lenguaje ampliando de esta manera su nivel de reconocimiento.

* Gabriel Cordero Sánchez publicó este trabajo en abril de 1985 en un congreso nacional. Lo inició en 1983 cuando era miembro de la Unidad de Investigación y Desarrollo del Departamento de Computación de la UPIICSA, profesionalmente desempeñaba el cargo de Subdirector de Análisis y Diseño de Sistemas en la Dirección General de Información Administrativa (DGIA) de la Secretaría de Educación Pública (SEP)

INTRODUCCIÓN

El documento consta de tres partes fundamentales; la primera incluye a manera de antecedente una breve reseña de la evolución de los sistemas de recuperación de datos (aunque es una realidad que el enfoque primario aún se usa) y se enuncia el objetivo del reconocedor junto con algunas perspectivas de su desarrollo. La segunda parte consta de una descripción del esquema conceptual (medio ambiente del sistema) y de los elementos del reconocedor, la cual define funcionalmente la operación del mismo dentro del subsistema de recuperación de datos. Por último se presenta el flujo de operación del reconocedor desde el punto de vista interno y da una idea más clara de la estructura de su construcción.

Se comenta la posibilidad no solo para la recuperación de datos, sino para la construcción de herramientas más avanzadas.

I ANTECEDENTES

Es evidente (para aquellos que estén relacionados) que ha existido un desarrollo importante en la metodología tanto del análisis y documentación de sistemas como para el diseño y construcción de programas. Lo anterior ha tenido como consecuencia, precisamente basada en la oportunidad de contar con nuevas y mejores herramientas; una evolución en los métodos para recuperación de datos, básicamente de los sistemas en los que están involucrados los datos (viviendo en algún dispositivo de almacenamiento) y los usuarios (que bajo cierto tratamiento —proceso— y a través de un medio, usualmente físico, requieran esos datos como parte de la información necesaria para su actividad dentro del sistema de información de la organización a la que pertenecen).

Una vez que la computadora pasó a ser una herramienta para las actividades operativas de la administración de cualquier organización, la necesidad de establecer un mecanismo de comunicación entre los requerimientos de un usuario y la posibilidad del proceso masivo de datos para su presentación (tradicionalmente en reportes) creció tal vez en mayor proporción

que el número de aplicaciones que podían introducirse a una computadora. Esto dio como resultado que los hasta ese momento los usuarios, trataran de entender el lenguaje (terminología) de los programadores y estos últimos trataran de conocer y representar la aplicación de aquellos. Surge una nueva disciplina, llamada, el análisis de sistemas administrativos, teniendo su origen en usuarios transformados en programadores y programadores transformados en usuarios. Hoy todavía, existen organizaciones que basan sus aplicaciones en este enfoque tradicional. En la figura 1 se muestra como fue la relación entre un usuario y sus aplicaciones en una computadora. Un usuario o grupo de usuarios, normalmente con el deseo de automatizar las actividades operativas y con manejo de grandes volúmenes de datos, a través de un interprete (comúnmente un programador o analista-programador y en el mejor de los casos un experto en sistemas administrativos favorablemente conocido como analista), cuya función además de decidir los datos necesarios para el usuario, era la de interpretar los requerimientos de este último y con esto elaborar un planteamiento del problema para de inmediato, dar paso al diseño y construcción de los programas indispensables en este caso. El usuario obtiene los reportes previa definición del programador, que contienen sus datos (del usuario) previa definición también del programador.

Con la introducción de metodologías para el desarrollo de sistemas y la natural evolución humana en tecnología, la recuperación de datos toma el sentido interactivo como medio principal para la obtención de información. En la figura 2 se puede ver como es el flujo de datos del sistema de recuperación bajo este enfoque, digamos más actual. Este tipo de recuperación de datos, surge del desarrollo de sistemas de aplicación de software a la solución de problemas en el manejo de datos de un usuario en particular. El usuario auxiliándose con el interprete (para este caso un analista de sistemas administrativos), logra definir sus requerimientos para el proceso y la presentación de sus datos.

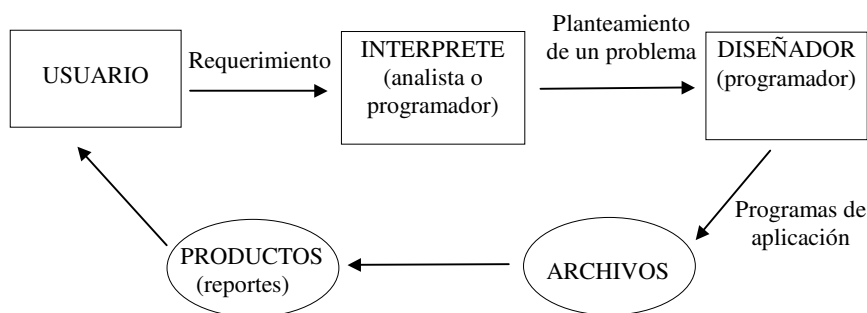


Figura 1. Recuperación de datos basada en el enfoque tradicional del desarrollo de aplicaciones

Estos requerimientos son planteados por el analista (para ese entonces un experto en las actividades del usuario) al diseñador de sistemas, quien dependiendo del enfoque del analista y del propio, procede con la definición del “sistema” para satisfacer las necesidades del usuario.

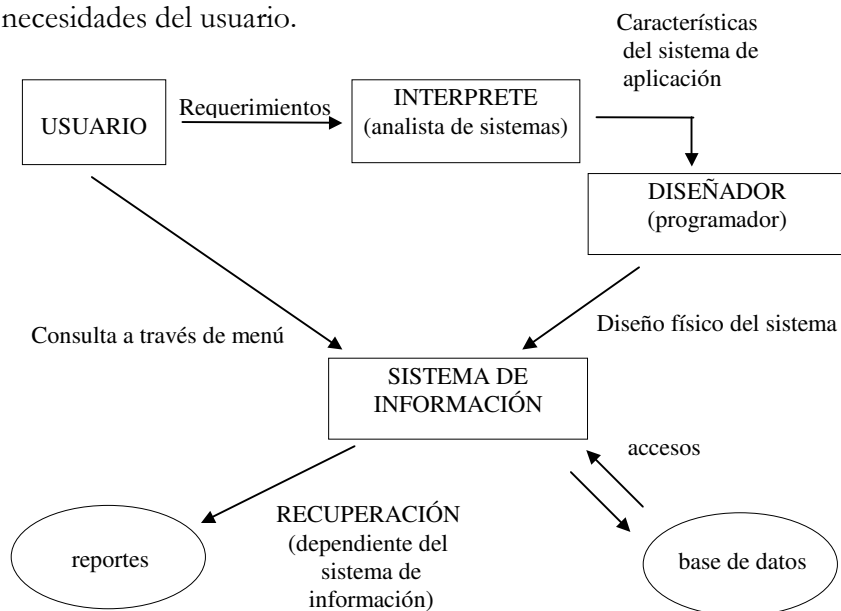


Figura 2. Recuperación de datos basada en el análisis de un sistema de información (enfoque interactivo)

Para este enfoque de la recuperación de datos, el diseñador de sistemas, haciendo uso de paquetes de software comercial define el sistema de aplicación en función de bases de datos exclusivas para un problema de un usuario en particular, y basa la operación de recuperación a través de un menú de opciones, Obligando con esto a que el usuario en principio, domine los términos del lenguaje en que fue definido el menú por el diseñador, *restringiendo además la posibilidad de la recuperación a sólo las opciones de ese menú*. Esto implica que un nuevo requerimiento del usuario tiene que ser diseñado y programado para su inclusión al sistema de aplicación (menú).

El estudio de organizaciones y su representación como sistemas de información ha venido ocasionando que la filosofía de construcción de software cambie hacia un enfoque menos dependiente de los sistemas de aplicación, conduciendo así al *diseño de software de uso general, que se distingue principalmente por la independencia entre el usuario* (como elemento del sistema de información de una organización), *y el sistema de aplicación*, diseñado específicamente como herramienta de apoyo a las actividades de ese usuario dentro de la organización. Este principio, el de *concebir los sistemas de aplicación como una herramienta y no como la solución*, hace necesario el contar con herramientas que permitan al usuario nombrar y tratar a los datos en términos del lenguaje del propio usuario, y no en los que defina el analista (característica de los dos enfoques anteriores), bajo el mismo principio es como fue construido el reconocedor descrito en este documento.

La figura 3 contiene una representación de lo descrito anteriormente, donde varios usuarios a través de un interprete, en este caso *software de uso general* y no una persona como se vio anteriormente, auxilian su actividad usando el mismo sistema de información soportado en una base de datos general. Por lo que el objetivo primario del reconocedor y de hecho su origen, es evitar que un usuario durante la recuperación de sus datos se vea en la necesidad de conocer términos de computación no indispensables para su propósito.

Este planteamiento deriva en varios objetivos de carácter secundario, ellos son:

- 1) *Facilitar para el usuario la recuperación de sus datos en términos de su propio lenguaje.*
- 2) *Eliminar la dependencia entre la recuperación de datos y los sistemas de información.*
- 3) *Construir una herramienta de uso común y general (no dependiente de algún sistema de información)*
- 4) *Construir una herramienta capaz de registrar nuevos términos de lenguaje de un usuario o de nuevos usuarios (aprendizaje)*

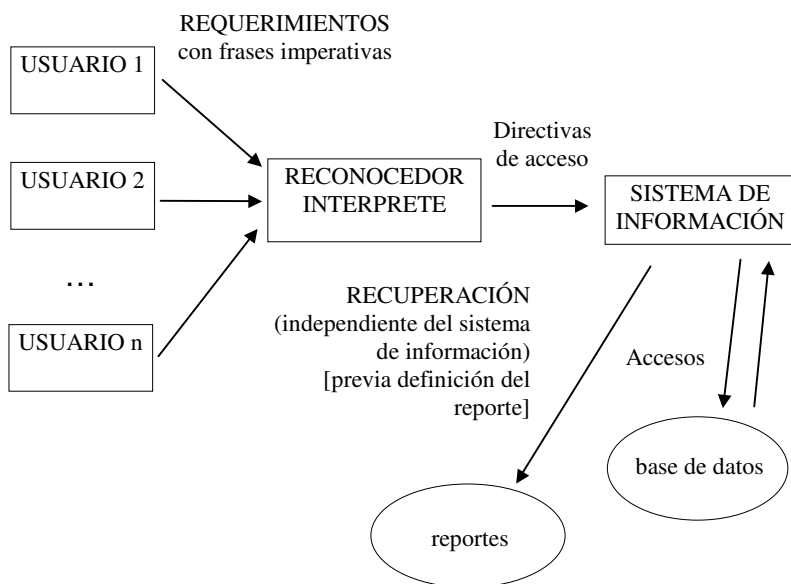


Figura 3. Recuperación de datos basada en el análisis y requerimientos del sistema de información de una organización
(construcción de herramientas de uso general)

II OPERACIÓN DE LA RECUPERACIÓN DE DATOS

La figura 4 muestra un ejemplo de la forma en que se usa el sistema de recuperación de datos tanto en forma interactiva (por terminal), como por reporte (previa definición). La operación se limita a la solicitud en forma imperativa de algunos datos por consultar, una condición es indicar el nombre del dato deseado dentro de la frase, para poder realizar la consulta (recuperación); Otro requisito necesario para efectos de reconocimiento es el de proporcionar al reconocedor entre // (diagonales) el valor de un dato del que se desee consultar sus características, por ejemplo:

Usuario: *Dame el artículo / 510101005/*

Sistema: *510101005, cinta magnética, 1,200 pies, pieza, ...*

Esta misma estructura y condiciones son las necesarias para la generación de reportes, que son distinguidos de la presentación en terminal desde el momento en que se inicia una sesión con el reconocedor.

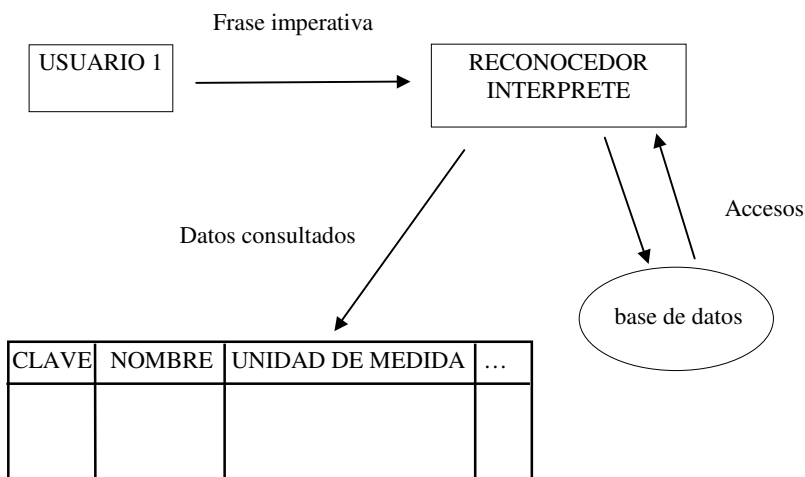


Figura 4. Operación de la recuperación

Las dos condiciones anteriores y las intrínsecamente relacionadas con la gramática incluyendo los elementos del lenguaje y el simple hecho de que el usuario conozca los datos (nombres de los datos) que maneja, constituyen los únicos requerimientos (de entrada) para que el sistema de recuperación pueda ser usado.

Otro punto muy importante y que forma parte del reconocedor es dar la facilidad al usuario para que registre nuevas palabras, es decir, que el reconocedor incremente su lenguaje y con esto su nivel de reconocimiento, las palabras que pueden registrarse en el catálogo de gramática pueden ser o no sinónimos de otras ya existentes. Un ejemplo de una sesión donde se tiene esto es:

Usuario: *Dame el objeto / 510101005/*

Sistema: *'objeto' no la tengo registrada, seleccione un sinónimo de entre: artículo, dependencia, pedido, presupuesto, saldo, ...*

Usuario: *artículo*

Sistema: *'objeto' es sinónimo de 'artículo', gracias.*

Sistema: *510101005, cinta magnética, 1,200 pies, pieza, ...*

De la misma manera, existe para cada solicitud de datos una forma de presentación. Esta forma de presentación está predefinida y registrada en los directorios correspondientes, (directorio de reportes, directorio de pantallas).

En el caso de la recuperación por pantalla, la presentación de los datos consultados puede ser de dos tipos. Una presentación hecha a través de formatos de pantalla predefinidos con una forma especial (pantalla fija), la otra presentación (de datos) se realiza por medio de una pantalla de formato general que se usa parcialmente condicionada por la misma solicitud (pantallas variables).

La predefinición de los formatos de presentación de datos (formatos de salida) es el único requerimiento (de salida) para una solicitud de consulta, interactiva o por reporte

III ESQUEMA CONCEPTUAL DEL SISTEMA DE RECUPERACIÓN DE DATOS UTILIZANDO UN RECONOCEDOR DE LENGUAJE NATURAL RESTRINGIDO

Como puede notarse hasta esta parte del documento; el *reconocedor como objeto principal de esta presentación, es solo un elemento más de lo que forma el sistema de recuperación de datos utilizando lenguaje natural restringido (SRD-LNR)*.

Como puede observarse en la figura 5, el SRD-LNR consta entre otros, de cuatro elementos, tal vez los más importantes: el primero de ellos es el usuario, quien finalmente es el afectado en forma favorable o desfavorable por la oportunidad, claridad y facilidad con la que pueda obtener información, que para efectos de esta aplicación en particular, están almacenados formando una base de datos. El(o los) usuario(s) usando una terminal, para afectos de la recuperación de datos, puede emitir tres clases de instrucciones (la mayoría de carácter imperativo y algunas asociativo):

- 1) *Solicitar la presentación* (consulta de algún dato o datos), para los cuales esta autorizado, y a través de la misma terminal.
- 2) *Definir un reporte* de sus datos o parte de ellos usando uno de los módulos del generador de reportes, en el que prácticamente la presentación del reporte la decide el usuario.
- 3) *Solicitar la emisión de un reporte* previamente definido indicando su referencia y el conjunto de datos que incluirá.

Las instrucciones 1) y 3), no importa de cual se trate, son recibidas directamente por el reconocedor, el segundo de los elementos del SRD-LNR, el reconocedor una vez identificada e interpretada la instrucción, basado en la gramática ya definida, tiene la función de invocar al generador de reportes o al módulo correspondiente del sistema de información, estos últimos, reciben, producto de la operación del reconocedor un conjunto de directivas de acceso, cuya ejecución da como resultado contar con los datos deseados listos para su presentación.

El sistema de información, tercer elemento del SRD-LNR, es el módulo encargado de realizar cualquier tipo de acceso a la base de datos a partir de las directivas generadas por el reconocedor. Adicionalmente auxiliándose por un manejador de pantallas y un directorio de datos, tiene la función de formatear los datos solicitados cuando estos sean presentados por la terminal. Este es uno de los caminos por el que el usuario recibe respuesta del SRD-LNR.

El generador de reportes, cuarto elemento del SRD-LNR y otro de los caminos por los cuales el usuario recibe respuesta a su solicitud, se auxilia del directorio de datos y basa su operación en las directivas emitidas por el reconocedor, en los accesos a datos por el sistema de información y en los reportes previamente definidos por el usuario y registrados en el directorio de reportes.

Todos los anteriores y algunos otros que serán descritos posteriormente, forman el medio ambiente del reconocedor y son parte del sistema de recuperación de datos.

1 EL RECONOCEDOR

En el punto anterior, se describió muy generalmente la función del reconocedor, y siendo éste el objeto del documento actual, será descrito a continuación en forma más detallada, conservando el marco trazado por el esquema anterior, y sin pretender con esto establecer la estructura interna y la construcción del reconocedor de no interés para los propósitos de esta presentación.

Tomando como referencia el diagrama de flujo de la figura 6, partimos nuevamente del elemento principal de la recuperación de datos, el usuario, que como ya se explicó solicita datos por medio de instrucciones de tipo imperativo y asociativas. Estas instrucciones son captadas en principio por el módulo de control

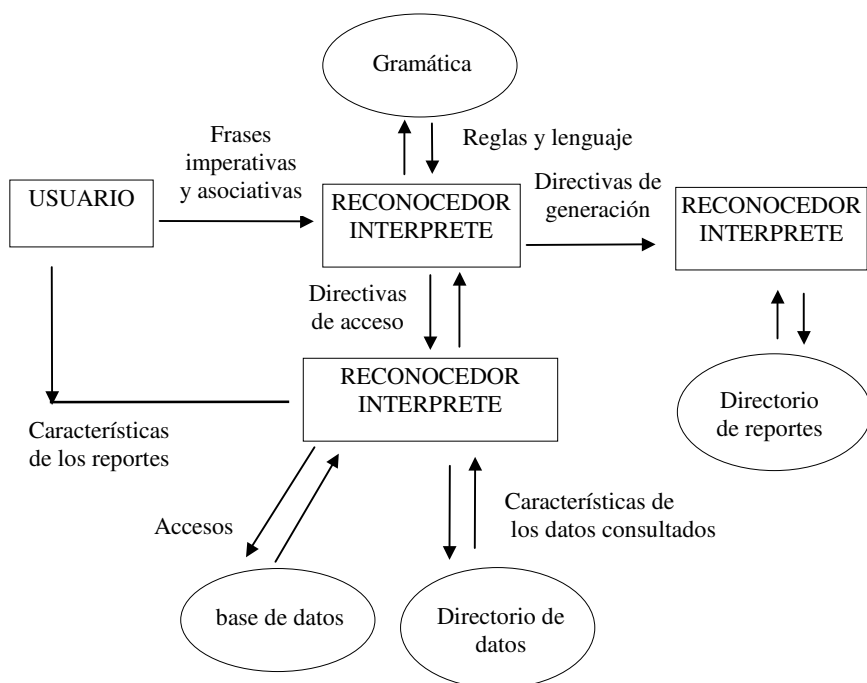


Figura 5. Elementos de Sistema de Recuperación de Datos utilizando Lenguaje Natural Restringido (SRD-LNR)

de accesos, que interactuando con el módulo reconocedor (no confundir con el reconocedor como elemento del sistema), tiene la función de verificar que (de acuerdo al tipo de consulta) el usuario esté autorizado para realizar la consulta de los datos solicitados. Los datos validos por usuario se encuentran registrados en el directorio de accesos. Una vez que un usuario ha sido identificado como válido, se hace el reconocimiento de la instrucción del usuario apoyado en la gramática previamente definida. Durante el reconocimiento de la instrucción pueden ser detectadas palabras desconocidas, las cuales, el módulo de ampliación del lenguaje con ayuda del usuario; puede asociarlas con algunas otras palabras ya conocidas o registrarlos como palabras nuevas, esto permite que el reconocedor aprenda a reconocer nuevas frases.

Quando una frase ha sido totalmente reconocida, el reconocedor codifica las instrucciones en el orden en que fueron encontradas

en la solicitud del usuario. Producto del reconocedor es una tabla de instrucciones ya codificadas y datos para las instrucciones.

Esta tabla es usada por el módulo interprete que en una primera fase hace una clasificación de las instrucciones basado en las características y en las relaciones de jerarquía y dependencia de los datos solicitados, ya clasificados genera una secuencia de directivas de acceso que pueden ser válidas tanto para el generador de reportes como para el módulo del sistema de información.

La segunda fase del interprete consiste en ejecutar las directivas de acceso generadas, por lo regular las que involucran al generador de reportes. El control de esta fase es compartido por el interprete y el sistema de información, si la consulta solicitada requiere de intervención del usuario. Por último, los datos solicitados, también soportados por el directorio de datos, son formateados con ayuda del manejador de pantallas, que permite que el usuario defina la forma en que deberán ser presentados en la pantalla de la terminal, de esta forma es como se establece el ciclo completo de la solicitud de consulta de un usuario hasta la presentación de los datos por la terminal.

El otro camino para obtener los datos, ya mencionado, que es a través de reportes y la forma en que opera el módulo del sistema de información, serán descritos de manera particular en los siguientes párrafos, aunque en forma breve por no ser el interés principal del presente.

2 EL GENERADOR DE REPORTES

En esta parte, el usuario puede tener comunicación con tres módulos principales del generador de reportes. El módulo de definición de reportes, a través del cual se puede definir la estructura características y contenido de un reporte.

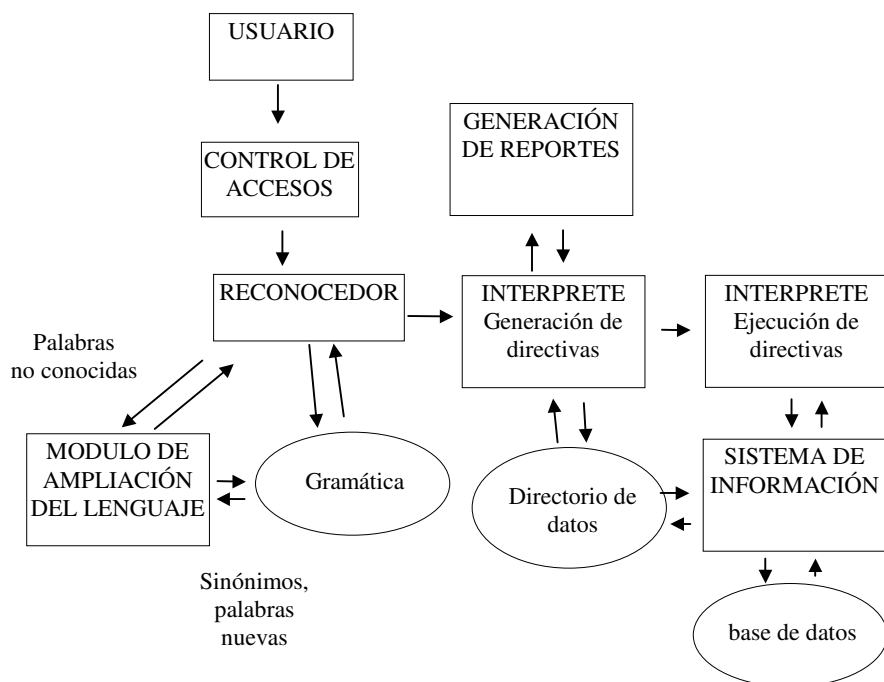


Figura 6. El Reconocedor

El reporte es registrado en el directorio de reportes previa validación de los datos del contenido contra el directorio de datos. La estructura considera el encabezado, cuerpo y pie del reporte en definición. Como características se entiende las condiciones de clasificación, salto y corte de control del reporte.

El módulo de definición de archivos, permite definir los archivos de datos a partir de los cuales se pueden obtener reportes, además de los que se obtienen a partir de la base de datos. Este módulo, aunque no representado en la figura, sirve de fuente de entrada para el directorio de datos.

Por último, el módulo de explotación o generación de reportes, cuya operación se soporta en reportes, archivos y datos ya definidos como se explicó, y se orienta por las directivas emitidas por el reconocedor con base en las instrucciones del usuario.

Cuando el reporte debe generarse a partir de la base de datos, el generador de reportes obtiene los datos a imprimir vía el módulo de control de accesos del sistema de información.

3 EL SISTEMA DE INFORMACIÓN

El sistema de información, elemento final de la recuperación de datos, tiene dos funciones básicas distribuidas en sus tres módulos principales. La primera, realiza los accesos a la base de datos con las directivas emitidas por el reconocedor. El módulo de control de accesos es encargado de esta función que conduce a dos posibles salidas, a través del generador la obtención de un reporte y el formateo de los datos accedidos, apoyado en el directorio de datos. El formateo de datos se hace de acuerdo a la forma de presentación requerida también previa definición usando el manejador de pantallas.

El último módulo, presentación de datos, se encarga de pintar los datos en el orden requerido en la pantalla de la terminal usada por el usuario.

Este conjunto de módulos fue identificado bajo el nombre de sistema de información por la razón de que es ésta la última parte de la recuperación y en la que están definidas las diferentes formas de presentación de datos, que en ultima instancia fueron solicitadas por el usuario y constituyen uno de los principales apoyos para la toma de decisiones y para guiar su actividad dentro de la organización.

4 FLUJO DE DATOS DURANTE LA RECUPERACIÓN

Tratando de dar una idea de la estructura interna y de las herramientas que fueron usadas para la construcción del reconocedor y del sistema de recuperación de datos en general, ha sido incluido este párrafo que describe el flujo de datos, desde la solicitud del usuario hasta la respuesta del sistema obtenida por éste.

La instrucción del usuario es descompuesta en palabras, cada una de ellas y su orden, es validada contra la gramática dando la opción a registrar (almacenar en la gramática), palabras no conocidas mediante *asociaciones y decisiones del usuario*. Una instrucción del usuario puede generar varias directivas de acceso, estas son colocadas en una cola en el orden en que fueron encontradas, simultáneamente asociado a la cola de directivas se genera una lista de los valores de datos que el usuario solicitó. Por medio de una lista bidireccional esta cola de directivas se clasifica, de acuerdo al orden de acceso de los datos solicitados, según el directorio de relaciones. Hecho lo anterior, se procede con la ejecución de las directivas, controlando con una pila (stack) aquellas directivas de ciclo, es decir aquellas que fueron originadas por una solicitud en la que el usuario desea obtener: datos, un conjunto de datos, un conjunto de datos de un conjunto de datos, etc.

La última fase tiene dos salidas, una usando los datos ya accedados, se presentan en un reporte, que como ya fue mencionado, debe estar definido. La otra salida se realiza a través de la pantalla de la terminal de usuario, misma por la que hizo su solicitud, esta parte se auxilia de un manejador de pantallas para efectos de formateo y presentación de los datos.

CONCLUSIONES

El reconocedor, como objeto principal de esta presentación, no significa por si mismo la creación de una herramienta sin igual, tecnológicamente hablando en lo que a computación se refiere; significa en cambio, la factibilidad, y realidad de hecho, del nivel de herramientas que pueden ser construidas y usadas, en un ambiente en el que posiblemente existan limitaciones materiales pero no de imaginación. La filosofía empleada en esta aplicación de la recuperación de datos y el reconocedor concretamente, establece las bases para el desarrollo y construcción de herramientas con mayor flexibilidad, orientadas hacia el usuario final.

Algunas de las posibles herramientas que pueden ser construidas basadas en el reconocedor son:

- 1) *Generadores de programas o programación en lenguaje natural*, estos solo con una ampliación al reconocedor para que sea capaz de interpretar instrucciones de tipo condicional.
- 2) *Bases de conocimiento* que podría también ser usadas para la recuperación de datos y reduciría en mucho la longitud de la instrucción de solicitud de consulta del usuario.
- 3) *Otra aplicación* y que ya se esta desarrollando es la de permitir a través del reconocedor no solo la recuperación de datos sino además la actualización, ambas con instrucciones de tipo imperativo.

Es importante, al menos por ser un hecho que pudo observarse con el uso de esta herramienta, hacer notar como la persona que opera con el reconocedor (y cabe mencionar que esta persona ha operado con otros sistemas de recuperación de datos), como realmente lo único que requiere es conocer la información que maneja en su trabajo, y al percatarse de no tener una vía preestablecida para obtener datos suele decidir por mejores opciones (presentaciones de datos) diferentes a las que de origen fueron planteadas por el usuario mismo. Esto tiene como consecuencia que un usuario vea y sienta al sistema como una verdadera herramienta y no como un algo normativo que lo eduque, pensamiento, bajo el cual aún se rigen muchos desarrolladores de sistemas.

REFERENCIAS

- [1] Knuth, Donald, "*The Art of Computer Programming*" vol. I y vol. III
- [2] Vetter, "*Data Base Design Methodolgy*"
- [3] Hopcroft, Aho, "*Introduction to Automata Theory and Languages*"
- [4] Warnier, "*Logical Construction of Programs*"
- [5] _____, "*SAASEP: Sistema de Apoyo Administrativo de la Secretaría de Educación Publica, Subsistema de Pago*"
- [6] _____, "*Feature Analysis of Generalized Database Managment Systems CODASYL Systems Comitee Technical Report*", mayo 1971