Multiple Fonts from a Single Font Optical Character Recognizer

Riccardo Ferrari, Malvina Duarte

Laboratorio de Electrónica Digital
Universidad Católica "Nuestra Señora de la Asunción"
Po. Box 1718
Asunción - Paraguay
email: rferrari@ledip.uucp, mduarte@ledip.uucp

Carlo Viscomi

Hypertex User Group
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Milano - Italia

Some available OCR systems provide documents with a single font, style and syze. The present work studies the way for producing documents with multiple fonts, style and size starting from a trainable, but single font, recognizer. The analysis we developped led to the definition a *recognizing methodology*. This methodology has been ipmlemented in a prototype that we describe in the paper. Finally we present an example of application to a Calculus textbook.

Algunos sistemas disponibles para la lectura óptica de documentos de papel dotados de reconocimiento de caracteres suministran documentos con un único tipo de carácter, estilo y cuerpo. En el presente trabajo se estudia como producir documentos con tipo de carácter, estilo y cuerpo múltiples, utilizando un reconocedor instruible de este tipo. El análisis del problema que hemos efectuado nos ha llevado a definir una *metodología de reconocimiento* que hemos implementado en el prototipo que presentamos. Al final de la presente nota se describe un ejemplo de aplicación de la metodología propuesta sobre un texto de análisis matemático.

Introducción

Para simplicidad de referencia, comenzamos introduciendo algunos términos. Con el término *símbolo* se entiende un elemento cualquiera del código ASCII y con *atributos* el tipo de carácter, de estilo y de cuerpo de un carácter.
La metodología de reconocimiento está estructurada en dos principales actividades. En la primera se efectúa un análisis preliminar de los símbolos a reconocer, de aqui se determina una clasificación en relación a los atributos de los cuales están dotados, se instruyen al reconocedor y al prototipo en base a la clasificación definida. La clasificación se basa en la agrupación de símbolos que presentan un atributo común. Luego se efectúa el reconocimiento con producción de documentos de tipo texto, cuyos símbolos presentan atributos comunes, es decir un único tipo de carácter, estilo y cuerpo. La segunda actividad consiste en la importación en el prototipo de estos documentos y en la conversión automática de los atributos de los símbolos sobre la base de la clasificación. El resultado que se obtiene es la reproducción en formato electrónico del texto sobre papel, fiel en todos los símbolos respecto a sus atributos originales.
El prototipo se presenta en su aspecto funcional y en las modalidades de uso de cada comando.
El artículo comienza con una breve descripción de los reconocedores de caracteres, clasificados en relación a funcionalidades y modalidades de uso.

1. Digitalización de un texto

1.1. Los reconocedores de caracteres

Los reconocedores (Optical Character Recognizer - OCR) se pueden dividir en tres grandes categorías: No entrenables, Entrenables, Automáticos.
Los OCR de tipo *No entrenable* se basan en las confrontaciones de conjuntos de modelos, llamadas clases de referencia. Tales conjuntos reproducen los símbolos a reconocer y son proveidos sólo para algunos atributos de los símbolos. En consecuencia, resultan reconocibles solo las impresiones que utilizan los mismos simbolos y solo con los atributos conocidos.
En los OCR de tipo *Entrenable*, las clases de referencia, llamadas "type table", son construidas por el usuario en un diálogo con el reconocedor el cual propone, una a la vez, áreas de bit-map que considera ser simples símbolos, en relación a las

cuales es posible introducir vía teclado el símbolo en código ASCII que la bit-map representa en forma gráfica. Tal posibilidad ha permitido gestionar los diferente atributos, como será detalladamente especificado a continuación.
Los OCR de tipo *Automático* están en grado de reconocer automáticamente, esto es sin ninguna fase preliminar de instrucción, textos caracterizados por una gran variedad de atributos de los símbolos, también en el interior del mismo documento.

1.2. Metodología de reconocimiento

El universo U del discurso es compuesto por todos los posibles símbolos s utilizables en cualquier documento de tipo textual. Para cada elemento de U se definen cuatro propriedades a(s), t(s), c(s), e(s), que describen respectivamente el valor de codigo ASCII[1] y los atributos: tipo (font), cuerpo (size) y estilo (style) del símbolo mismo. Se consideran símbolos diferentes aquellos que difieran por una o más de las propiedades.
Los símbolos que tienen mismo código ASCII constituyen las clases de una relación de equivalencia sobre U. Indicamos con ASCII* el conjunto de todas las clases de ASCII-equivalencia.
Considerado un documento de tipo texto (indicado con t), que consiste en una particular secuencia de símbolos, denotamos con S el conjunto de los símbolos que aparecen en t. Se definen atributos del documento las propiedades tipo, estilo y cuerpo (indicados respectivamente con t(t), c(t) y c(t)) de la mayoria de sus símbolos. En la operación de asignación inicial se indican al reconocedor los valores de los atributos del documento. Estos atributos serán asignados a todos los símbolos del documento reconocido.
Con el término símbolo general se entiende un símbolo cuyos atributos coinciden con aquellos del documento al que pertenece. Se define conjunto de los Símbolos Generales SG:

$$S_G = \{ s \mid s \in S, (t(s) = t(t)) \cap (c(s) = c(t)) \cap (e(s) = e(t)) \}.$$

Con símbolo especial se identifica un símbolo que posee por lo menos un atributo diferente de los atributos del documento al cual pertenece. El conjunto de los Símbolos Especiales SE está definido como:

$$S_E = \{ s \mid s \in S, (t(s) \neq t(t)) \cup (c(s) \neq c(t)) \cup (e(s) \neq e(t)) \}.$$

---

1 Consideramos aqui el código ASCII en versión extendida, osea con 256 caracteres

En la Fig. 1.1 se muestra un ejemplo de símbolo especial:

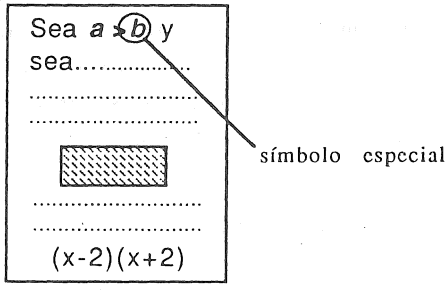

Fig 1.1: Ejemplo de símbolo especial

Con símbolo de pasaje se entiende una secuencia de dos símbolos, de los cuales el primero es el representante de una clase de equivalencia ASCII que no aparece en el texto t y cuyos atributos sean los del texto, y el segundo es un símbolo general. El conjunto de los Símbolos de Pasaje S_P resulta así definido:

$$S_P = \{pg \mid p \in (U\text{-}S_G),\ g \in S_G,\ [(t(p) = t(t)) \cap (c(p) = c(t)) \cap (c(p) = c(t))]\}.$$

Podemos elegir p entre todos los elementos de $(U\text{-}S_G)$ que además compartan los atributos del texto, entonces tenemos $(|ASCII^*| - |S_G|)$ diferentes candidatos a ser posibles p.
Resulta que si consideramos la cardinalidad del conjunto S_P, ésta es dada entonces por:

$$|S_P| = (|ASCII^*| - |S_G|) \cdot |S_G|$$

y representa la cota máxima de símbolos de pasaje y, como veremos a continuación, también de símbolos especiales manejables por el sistema.
La metodología de reconocimiento definida está constituída por la siguiente secuencia de fases:

1) búsqueda en el texto sobre papel de todos los símbolos especiales y creación del "Conjunto Especial" cuyos elementos a la vez son conjuntos. En efecto el "Conjunto Especial" C_E está definido como:

$$C_E = \{ C_{E1}, C_{E2}, \ldots, C_{En} \}$$

donde n es el número de conjuntos especiales encontrados en un documento, y donde el Conjunto Especial k-ésimo C_{Ek} es el conjunto cuyos elementos son símbolos especiales caracterizados por diferentes valores ASCII e iguales atributos de tipo, estilo y cuerpo.
En la figura 1.2 está esquematizado el paso 1.

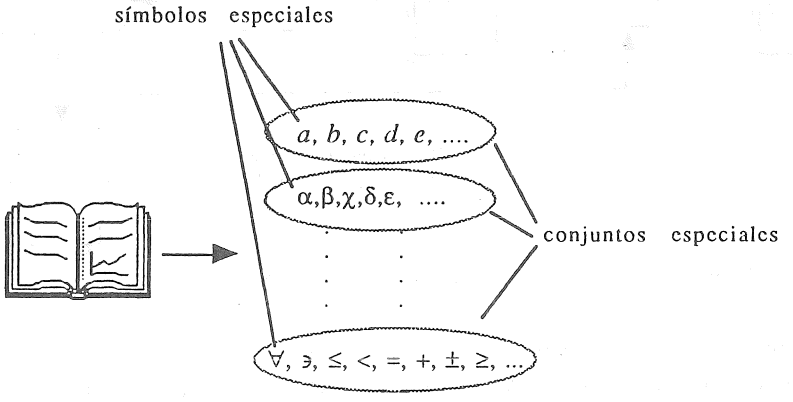

Fig. 1.2: Identificación de los símbolos especiales y definición de los Conjuntos Especiales

2) para cada símbolo especial de cada conjunto especial se procede a la creación de las asociaciones "*símbolo especial → símbolo de pasaje*" como es mostrado en la figura 1.3. Estas asociaciones están definidas por la función biunívoca *Asociación* A:

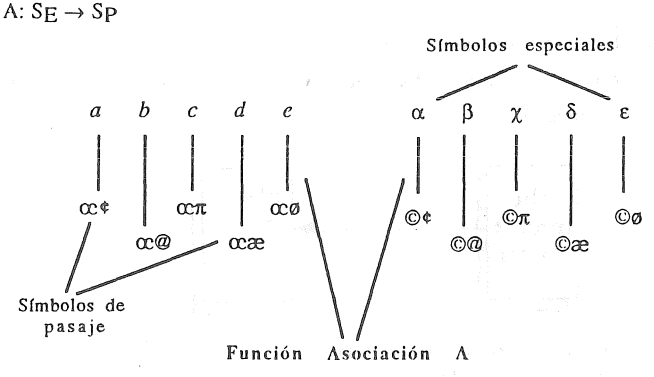

Fig. 1.3: Definición de las asociaciones "símbolo especial Æ símbolo de pasaje" y de la función "Asociación" Λ

3) instrucción de un OCR de tipo entrenable con creación de una *type table* sobre los atributos de los símbolos generales y especiales, utilizando la función Asociación antes definida; reconocimiento del texto (fig. 1.4).

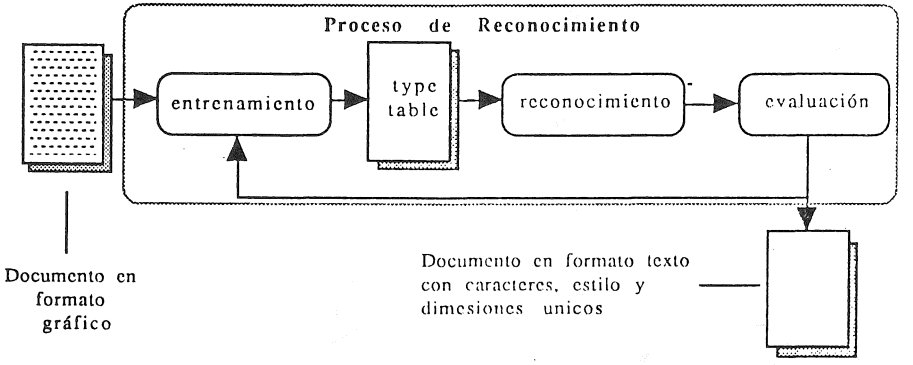

Fig. 1.4: Reconocimiento de texto con OCR de tipo entrenable

4) instrucción del instrumento sobre las asociaciones anteriormente definidas en sentido inverso (*símbolo de pasaje* → *símbolo especial*), definiendo la Asociación Inversa:

$$A^{-1} = (A)^{-1}$$

que es la función inversa de la función Asociación, obviamente también biunívoca:

$$A^{-1}: S_P \rightarrow S_E$$

5) sustitución global de los símbolos de pasaje en los respectivos símbolos especiales utilizando la función Asociación Inversa. El resultado final está representado en la figura 1.5:

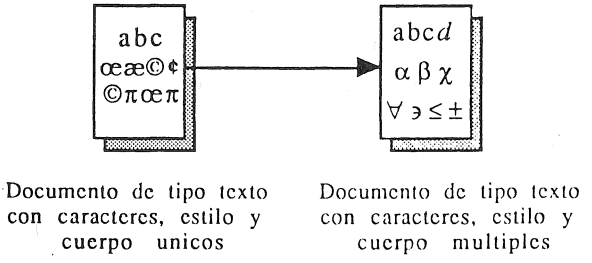

Fig. 1.5: Conversión símbolo de pasaje → símbolo especial

Las fases hasta ahora descritas pueden ser esquemáticamente resumidas a continuación, con la introducción de algunas ulteriores definiciones e introduciendo el operador *clausura* (*), que indica la concatenación de una secuencia, nula, finita o infinita, de símbolos pertenecientes al mismo conjunto al cual pertenece el símbolo que tiene por argumento.

texto inicial t_i, texto en formato de papel impreso constituido por una secuencia de símbolos generales y de símbolos especiales:

$$t_i = \{ (s)^* \mid s \in S_G \cup S_E \}; S_E$$

texto de pasaje t_p, texto en formato electrónico constituído por la secuencia de símbolos generales y de símbolos de pasaje:

$$t_p = \{ (s)^* \mid s \in S_G \cup S_E \};$$

texto final t_f, texto en formato electrónico, constituído por la secuencia de símbolos generales y de símbolos especiales:

$$t_f = \{ (s)^* \mid s \in S_G \cup S_E \};$$

Transformación Inicial T_I, proceso de reconocimiento, guiado por la metodología de reconocimiento , que a partir de un texto inicial y de los conjuntos asociaciones ya definidas, produce el correspondiente texto de pasaje:

$$T_I : t_i \times A \rightarrow t_p;$$

Transformación Final T_F, proceso que a partir de un texto de pasaje y de los conjuntos de las asociaciones inversas encima definidas, obtiene el correspondiente texto final:

$$T_F : t_p \times A^{-1} \rightarrow t_f.$$

El esquema de flujo de la metodología general está dado en la fig. 1.6:

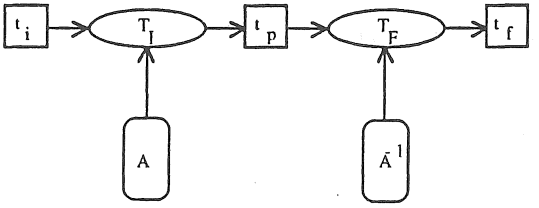

Fig. 1.6: metodología de reconocimiento

2. La estructura funcional del instrumento

La estructura funcional de base del instrumento es dada por la siguiente figura:

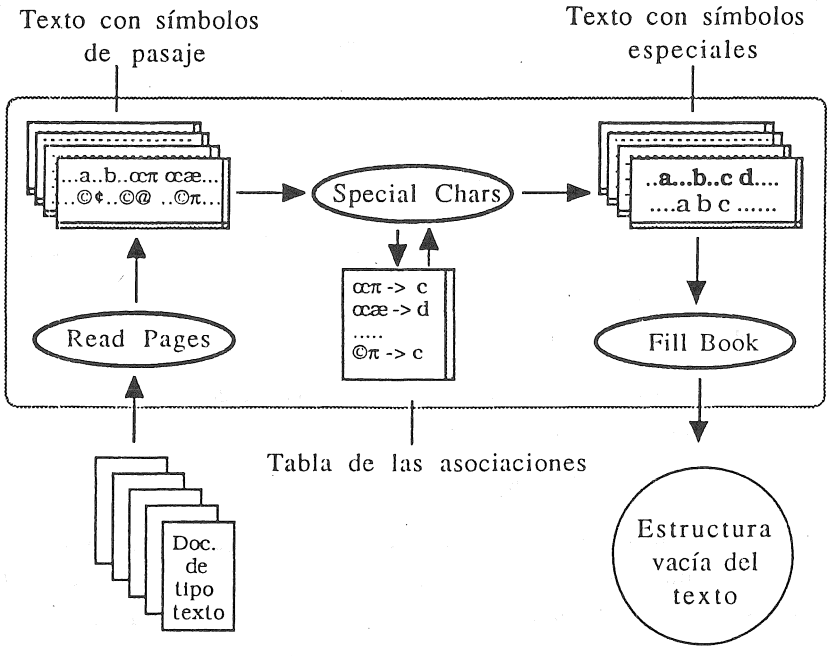

Fig. 2.1: Estructura funcional básica del sistema

El diseño muestra las funcionalidades principales y el tipo de documento en ingreso y en salida del sistema.

Las funcionalidades pueden ser reagrupadas según dos principales actividades: importación de documentos de tipo texto en el sistema (función "Read Pages"); definición e inserción de la función Asociación Inversa con construcción de las Tablas de asociaciones y sustituciones de los símbolos de pasaje en los símbolos especiales correspondientes (función "Special Chars"); exportación de los documentos en una estructura vacía para la realización de un texto electrónico (función "Fill Book").

La fig. 2.1 muestra la secuencia de operaciones de una sesión completa de trabajo; no está definido un orden de ejecución temporal entre las distintas fases,

mientras existe una serie de controles menores sobre el uso correcto del sistema por parte del usuario.

3. Modalidad de uso del tool.

3.1. Operaciones generales

Es necesario especificar al inicio de cada sesión de trabajo el *pathname* de la carpeta de la cual leer los documentos de tipo texto, por el contrario asume el valor de default que corresponde a la última especificación dada.
Es posible escoger el atributo carácter para los documentos presentes en el sistema, que volverá a ser en consecuencia el carácter para el texto general. Tal funcionalidad se puede llamar en cada instante durante la sesión de trabajo, teniendo presente, sin embargo, que la elección, o el cambio de carácter conlleva la pérdida de eventuales símbolos especiales ya restablecidos, puesto que la nueva elección es aplicada sobre el texto de cada documento presente en el sistema.

3.2. Lectura de los documentos (Read Pages)

Un documento de tipo texto, a fin que pueda ser importado en el sistema, debe tener por nombre un número. La elección ha sido dictada por la comodidad del uso, en el caso se tratan de reconocer un texto con numerosas páginas, ya que el instrumento no requiere el *pathname* de los documentos particulares sino más bien informaciones de carácter más general. En efecto la carga de los documentos es hecha por grupos de páginas, indicando la primera y última página del grupo mismo. El grupo, en el interior del sistema, puede ser indentificado con un capítulo nuevo (*Capitolo Nuovo*), en ese caso es necesario indicar también el nombre del capítulo, o bien puede ser insertado en el interior de un capítulo existente (*Nuova Pagina...*) después de la página corriente. Es posible remover una página (*Cancella Pagina...*), o bien un capítulo (*Cancella*), por vez.
La navegación entre los documentos leídos es de tipo secuencial, función permitida a travéz de una *window* de tipo "palette". Es posible navegar a nivel de capítulos, o de páginas en el interior del mismo capítulo, por un elemento (página o capítulo) al precedente o sucesivo, al primero o al último. Están presentes indicaciones sobre la página y capítulo corriente para dar una mínima orientación al usuario.

Finalmente está presente una función que permite de posicionarse directamente sobre una página (*Go to...*).

Las funcionalidades descritas están representadas en la figura 3.2.1.

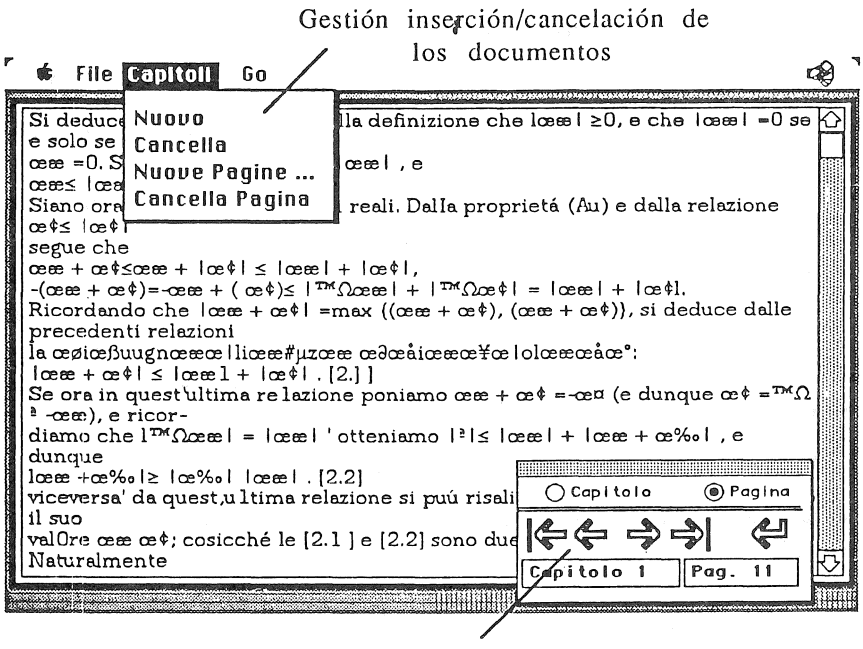

Fig. 3.2.1: La lectura y la navegación de los documentos

3.3. El tratamiento de los caracteres especiales (Special Chars)

Las funcionalidades disponibles en el tool permiten de insertar las asociaciones símbolos de pasaje->símbolos especiales y de aplicar las conversiones en base a los mismos en los documentos presentes en el sistema.

Han sido considerados 46 símbolos especiales tomados del código ASCII extendido, con carácter "Bookman", estilo normal y cuerpo 12, y a cada uno de los cuales viene reservada una "Tablas de Asociaciones" para las asociaciones pertenecientes al grupo correspondiente.

En tales hipótesis el número de asociaciones máximas obtenibles está dado por 210 (o sea 256, la cardinalidad del código ASCII extendido, sustraído de 46, reservados como símbolos de pasaje), multiplicado por 46, el número máximo de símbolos iniciales establecidos para los símbolos de pasaje. El resultado resulta ser 9660 posibles asociaciones, o sea 9660 símbolos especiales gestionables por el sistema. La experiencia muestra que éste es un número más que suficiente para tratar las partes textuales de textos también científicos, normalmente caracterizados por una vasta heterogeneidad de atributos. La figura 3.3.1 muestra la "Symbols Table", o sea la tabla índice de los símbolos especiales, que permite la selección de uno de ellos a fin de posicionarse en la tabla de inserción que se le ha asociado.

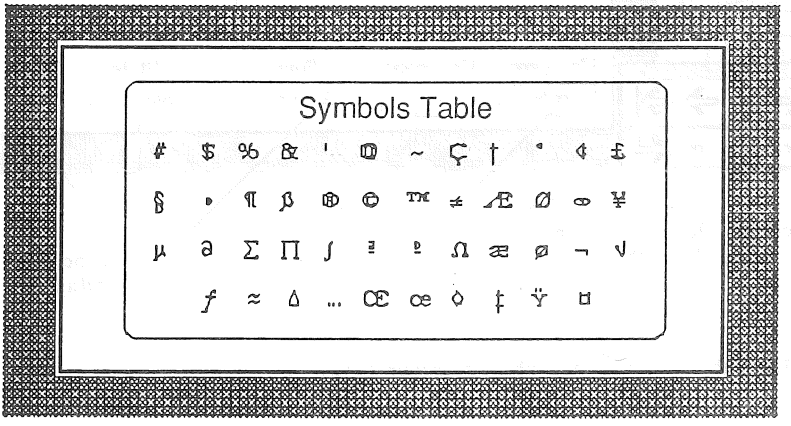

Fig. 3.3.1 : Tabla generál de los 46 símbolos especiales con atributo de caracter "Bookman"

La figura 3.3.2 presenta la tabla correspondiente al símbolo de pasaje inicial "œ".

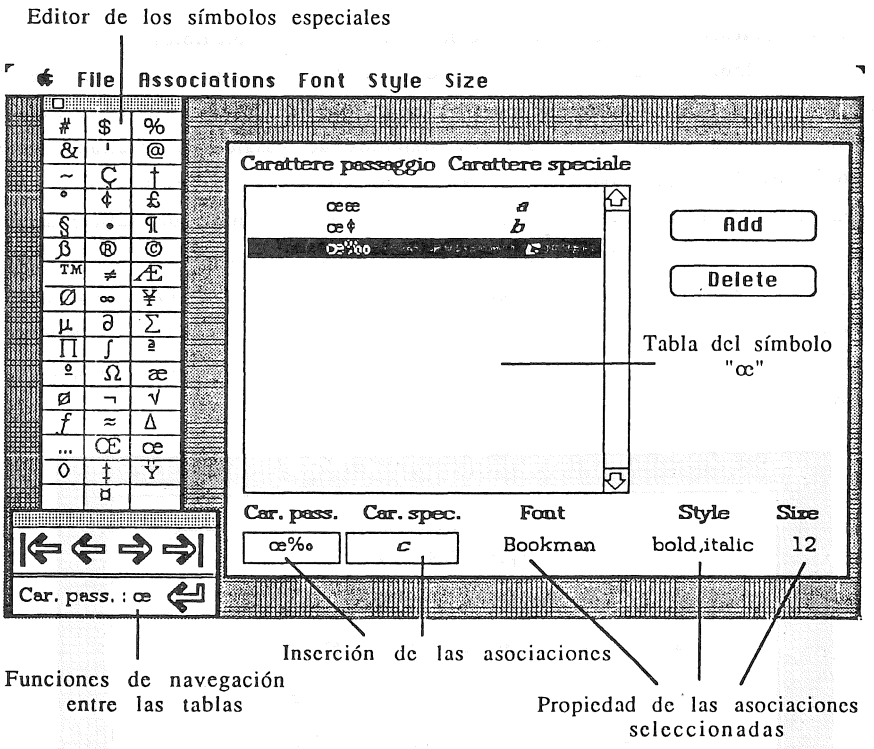

Fig. 3.3.2: Tabla asociadas al símbolo de pasaje inicial "œ".

La inserción de las asociaciones (*Add*) es a través del teclado con el auxilio de un Editor para los símbolos especiales (*AssociationEditor*), a fin de simplificar y volver más rápidas las operaciones; las cancelaciones de las asociaciones está permitida para aquella actualmente seleccionada (*Delete*) o bien para todas las presentes en el sistema (*Delete All*).

Es posible moverse entre las tablas secuencialmente, de la presente a la sucesiva o precedente, o bien a la primera y última del sistema, y finalmente jerárquicamente llamando nuevamente la Symbols Table (*Association Table*), y seleccionando el símbolo de pasaje inicial al cual está asociada la tabla buscada.

Para la fase de sustitución de los símbolos de pasaje en los símbolos especiales, está disponible una serie de funcionalidades que permite de operar sobre la asociación seleccionada (*Apply Selection*), sobre todas las asociaciones de la tabla corriente (*Apply Group*), sobre todas las asociaciones de todas las tablas (*Apply*

All). En cada caso, las sustituciones se realizan sobre todos los documentos presentes en el sistema. La figura 3.3.3 nos muestra las alternativas funcionales descritas:

Fig. 3.3.3: Funciones para la gestión de las asociaciones

Al término de las sustituciones, se suministran algunos datos sobre el número de sustituciones efectuadas por cada asociación, así como sobre el sistema general. La fig. 3.3.4 muestra un ejemplo de sustitución sobre un grupo de 10 páginas importadas en el sistema.

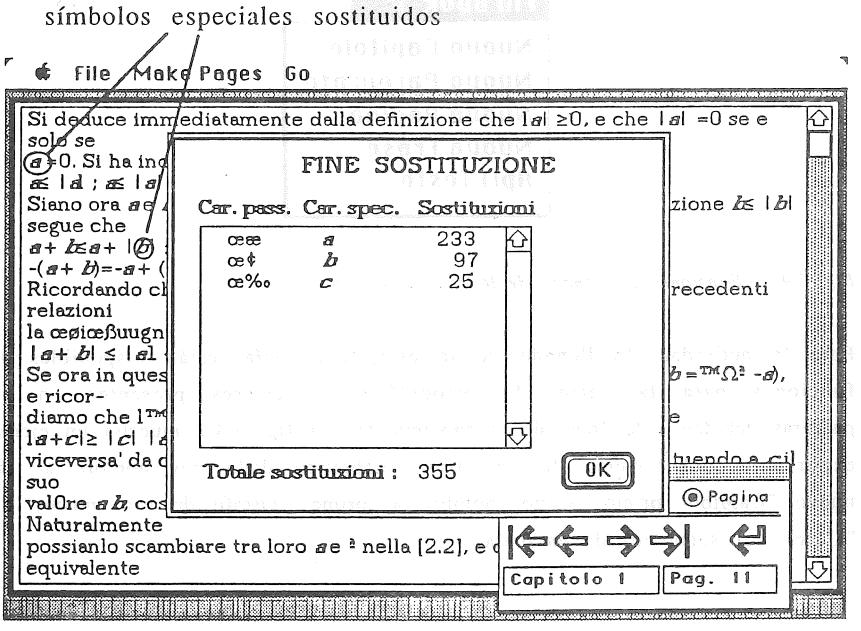

Fig. 3.3.4: Informaciones estadísticas sobre operaciones de sustituciones

3.4. Producción de un texto electrónico (Fill Book)

Un ejemplo del uso final del sistema está dado por la creación del texto electrónico sobre la base de los documentos reconocidos.
En el instrumento se cuenta con funciones que permiten el transporte automático de los documentos presentes en el sistema, en la predispuesta estructura vacía de un texto electrónico.
El texto seleccionado en el documento corriente por el autor puede ser transportado como frase (*Nuova Frase*). La operación de exportación de una frase puede ser precedida por instrucciones para la creación en el texto de un nuevo capítulo (*Nuovo Capitolo*), de un nuevo parágrafo (*Nuovo Paragrafo*) o de una nueva página (*Nuova Pagina*): en los dos primeros casos es posible especificar el nombre del capítulo y del parágrafo que se utilizarán para la orientación del lector del texto. Diversamente, el copiado de una frase se desarrolla según la modalidad de la última efectuada. Además es posible visualizar el estado actual del texto electrónico pidiendo de ingresar en simple lectura (*Apri Testo*).
La fig. 3.4.1 nos muestra las alternativas de la funcionalidad descrita.

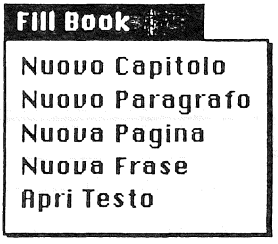

Fig. 3.4.1: Funciones de llenado de la estructura vacía

Para la actividad de llenado de la estructura vacía, están disponibles algunas funciones para la corrección ortográfica de errores presentes en algunas palabras debidas a la frase de reconocimiento. La fig. 3.4.2 muestra un ejemplo de exportación del sistema de una frase, habiendo dado por nombre al capítulo nuevo "Capitolo primo" y por nombre al primer párrafo de este mismo capítulo "L'area del segmento di parabola".

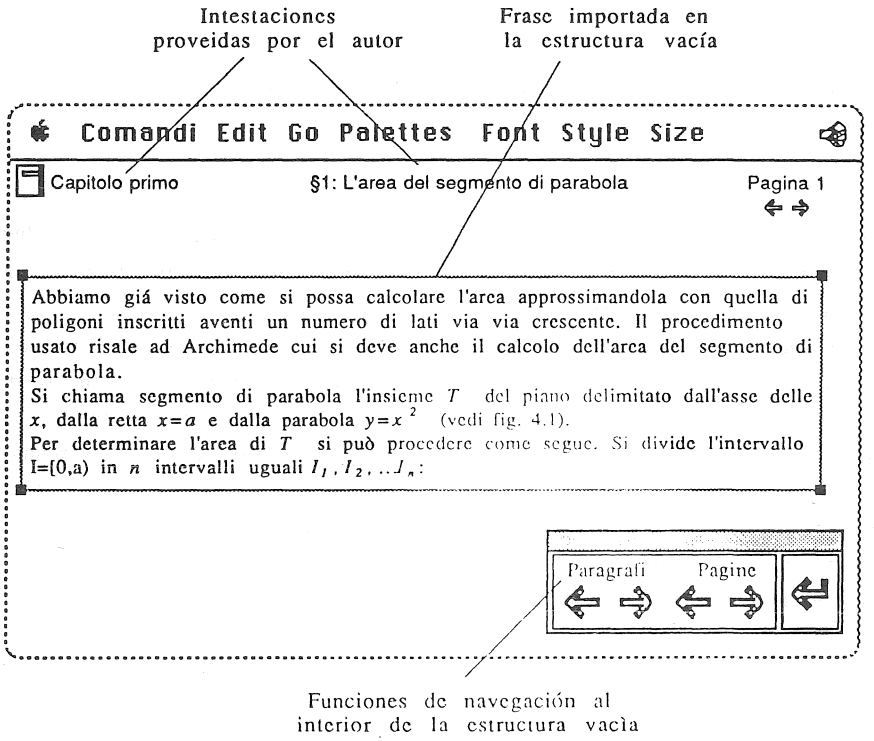

Fig. 3.4.2: Ejemplo de importaciòn de una frase

4. Una experiencia de digitalización

Ha sido conducida una experiencia de digitalización sobre el texto "Analisi Matematica I", autor Enrico Giusti, adoptado en el curso de Doctorado de Ciencia de la Información de la Universidad de Milán. El texto se ha prestado en modo particular al análisis del problema del "data-entry" puesto que presenta una vasta heterogeneidad de atributos, caracteres, estilos, cuerpos, además en número elevado.

Durante el proceso de reconocimiento se ha construído una *type table* que presenta un conjunto de 256 bip-map diversas, de los cuales 107 símbolos de pasaje; por cada símbolo a reconocer han sido hechas más de una entrada, o bien más relaciones bit-map de un mismo símbolo; de 15 normalmente a 40, a fin de aumentar el nivel de reconocimiento. Se han contado así en total 1122 entradas.

En relación a los tiempos de instrucción del reconocedor, ha sido necesario cerca de cuatro días de trabajo, mientras que para el reconocimiento se han estimado

valores de 2 minutos por página normal (en el orden de 1500 caracteres) hasta 4 minutos por página de alta densidad de texto, aquellas del Giusti que contienen las noticias históricas (3400 caracteres). El reconocimiento se realiza en modalidad *batch*: se han soportado sesiones de trabajo con sesenta documentos en entrada sin problemas.

Es de notable importancia construir *type tables* completas y precisas, que trabajen con un nivel alto de reconocimiento, puesto que así se reduce drásticamente la intervención manual de corrección de los errores que comunmente un OCR genera.

Una vez reconocidos todos los 289 documentos, se ha utilizado el prototipo presentado en el párrafo anterior: se han leído los documentos en tiempos del orden de la decena de minutos; se han insertado los grupos de asociaciones en las respectivas tablas de inserción y se ha operado la conversión de las asociaciones. Esta última actividad ha requerido tiempos de ejecución en el orden de 15 horas para un total de cerca de 45.000 sustituciones, calculando tiempos medios de 1,2 segundos por sustitución. En la figura 4.1 está representado un ejemplo de expresión matemática reconocida.

$$\varphi = 2\varphi_{[-1,0)} + 2\varphi_{[0,1)} - \varphi_{[1,2)}$$

Fig. 4.1: Ejemplo de expresión matemática reconocida

Agradecimientos

El presente trabajo ha sido realizado en el ámbito de la cooperación entre el Laboratorio de Electronica Digital (LED) de la Universidad Católica "Nuestra Señora de la Asunción" y el Hypertex User Group (HUG) del Departamento de Ciencia de la Información de la Universidad de los Estudios de Milan, dentro del proyecto "Personal University".

Los autores agradecen al Ministerio de la Universidad e Investigación Científica de Italia y a la Universidad Católica "Nuestra Señora de la Asunción" del Paraguay por los medios puestos a disposición.

Se agradece a Gianni Degli Antoni y Francesca Greselin por las valiosas sugerencias y a todos los miembros del LED y del HUG por la colaboración.